Dissertation

submitted to the

Combined Faculties of the Natural Sciences and Mathematics

of the Ruperto-Carola-University of Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

Put forward by

Sven Dennis Kügler

born in: Bad Salzungen

Oral examination: December 9th, 2015

On the application of machine learning approaches in astronomy: Exploring novel representations of high-dimensional and complex astronomical data

Referees: Prof. Dr. Jochen Heidt
Prof. Dr. Joachim Wambsganß

To my family.

Zusammenfassung

Das Ziel der vorgestellten Arbeit ist die Anwendung datengetriebener Methoden auf komplexen und hoch-dimensionalen astronomischen Datenbanken. Der Schwerpunkt der Arbeit liegt dabei in der Erforschung neuer Datenrepräsentationen, um die Analyse der Daten mit Hilfe existierender Methoden des statistischen Lernen zu ermöglichen. Anhand von verschiedenen wissenschaftlichen Anwendungen werden die Vorteile der untersuchten Ansätze für Klassifizierungs-, Visualisierungs- und Regressionsaufgaben an astronomischen Daten aufgezeigt.

Im ersten Teil der Arbeit wird eine alternative Methode zur Bestimmung von spektralen Rotverschiebungen vorgeschlagen, welche die, von SDSS bestimmten, Rotverschiebungen als Wissensbasis nutzt. Die neue Darstellungsweise der Daten enthält hierbei nur Informationen, welche für die Bestimmung der Rotverschiebung, sowie der Detektion von multiplen Rotverschiebungen, notwendig sind. Anschließend wird eine neuartige Repräsentation von regelmäßigen Zeitreihen vorgestellt, basierend auf wiederkehrenden neuronalen Netzen. Dies erlaubt eine explorative Untersuchung von großen nicht-klassifizierten Datenbanken. Danach wird die Verwendung von Gaußschen Mischverteilungsmodellen als Darstellung für den statischen Teil von unregelmäßigen Zeitreihen diskutiert. Diese Darstellung ist allgemeiner formuliert als die Darstellung durch einzelne Merkmale, da sie die Einbeziehung photometrischer Unsicherheiten ermöglicht und nicht durch systematische Beobachtungseffekte beeinflusst wird.

Abstract

The goal of the presented work is the application of data-driven methods on complex and high-dimensional astronomical databases. The focus of the work is the exploration of novel data representations in order to enable the use of statistical learning approaches in the analysis of data. With the help of diverse science cases, the advantages of the introduced approaches for classication, visualization and regression tasks are shown by applying the developed methodology to astronomical data.

In the first part, an alternative approach for estimating redshifts of spectra by using the knowledge about the redshifts provided by the SDSS pipeline is presented. A novel data representation is employed which contains only information relevant for estimating the redshift and the detection of multiple redshift systems. Subsequently, a novel data representation for regularly sampled light curves based on recurrent networks is presented. This allows an explorative investigation of huge databases with unlabeled data. Finally, a new way of representing the static part of irregularly sampled light curves by a mixture of Gaussians is discussed. This representation is more general than the extraction of features, as it allows the inclusion of photometric uncertainties and avoids the introduction of observational biases.

Contents

Acronyms

AE	Autoencoder
ANN	Artificial neural network
ASAS	All Sky Automated Survey
ESN	Echo-state network
k**NN**	k nearest neighbors
LMC	Large Magellanic cloud
MAD	Median absolute deviation
OGLE	Optical Gravitational Lensing Experiment
PCA	Principal component analysis
SDSS	Sloan Digital Sky Survey
SMBHB	Super-massive black hole binary
SMC	Small Magellanic cloud
SOM	Self-organizing map
SVM	Support vector machine

List of Figures

List of Tables

Chapter 1

Introduction

With the advances in instrumentation, data acquisition and data processing, the amount of publicly available astronomical data increased enormously. This technical development increased the observational efficiency of sky surveys and thus the produced amount of data. Famous examples are SDSS (Sloan Digital Sky Survey, e.g., SDSS DR10, Ahn *et al.*, 2014) in the optical or FIRST (Faint Images of the Radio Sky at Twenty-Centimeters, Becker *et al.*, 1995) in the radio bands. Alongside imaging and spectroscopic surveys, another rapidly growing area of astronomical databases emerged, the time-domain astronomy. Examples are CRTS (Catalina Real Time Survey, Drake *et al.*, 2009) or the Kepler Space Mission (Borucki *et al.*, 2010) that have studied the temporal behavior of millions of objects. With future surveys, such as the LSST (Large Synoptical Survey Telescope, Ivezi *et al.*, 2011), the data produced per night will easily exceed the terabyte scale and will create its own challenges in terms of data processing and analysis. The open-access data policy in astronomy makes it an interesting field for the application of big-data methodology for both computer scientists and statisticians alike (e.g., Tiňo and Raychaudhury, 2012). Besides that, huge efforts have been made by the virtual observatory community (e.g., GAVO[1] and IVOA[2]) to allow a fast and easy access to the large variety of astronomical data. The two most outstanding and most widely used tools provided by the VO are the catalog access and matching facility *TOPCAT* (Taylor, 2005) and *ALADIN* (Bonnarel *et al.*, 2000), a tool for visualizing and matching astronomical imaging data.

In contrast to the exponentially growing amount of data, the number of approaches dealing with those high-dimensional, complex, possibly structured and partially irregular observations has only grown slowly over time (Ball and Robert, 2010). A specifically challenging problem in astronomy is the absence of a ground truth for most of the data and the missing possibility of designing own experiments. Additionally, the observations are often irregular, incomplete, and biased in the way they have been observed. Despite that, the observed relations are usually superpositions of several wanted and unwanted physical effects and their disentanglement is not always easily possible. As a consequence, astronomy requires high-quality analysis tools to gain knowledge from databases. For example, Graham *et al.* (2015) claim in a recent publication to have found a high-fidelity super-massive black hole binary candidate. The object was detected because out of $\approx 250,000$ spectroscopically confirmed quasars, it was the only one showing periodic behavior with a period of $P \approx 5.16\,\mathrm{yrs}$ in its light curve. This demonstrates impressively which efforts have to be undertaken to detect such an extraordinary behavior in huge databases. In recent literature, data-driven methodology has gained attention in the astronomical community. Here, the advantages of data-driven methodology for visualization and classification tasks (e.g., Matijevič *et al.*, 2012) as well as the detection of outliers (e.g., Protopapas *et al.*, 2006)

[1] German Astrophysical Virtual Observatory http://www.g-vo.org/
[2] International Virtual Observatory Alliance http://www.ivoa.net/

has been discussed. Machine learning methodology in astronomy has so far mainly focused on photometric redshift regression (e.g., Polsterer *et al.*, 2013), the classification of light curves (e.g., Richards *et al.*, 2011), and estimating periodicity in irregularly sampled observations (e.g., Graham *et al.*, 2013). Usually, learning tasks are not performed on raw observations but rather on extracted features, such as photometric measurements. This bears the risk that the choice of features has a large impact on the quality of the respective approach.

This thesis embeds the analysis of high-dimensional and complex data into the context of modern machine learning techniques. The focus of this work is to find more general data representations than features that are physically more meaningful. Consequently, the results are also more easily understandable and can provide a detailed insight into the drawbacks of existing methodologies. Apart from the development of a representation, the new approaches are applied to a broad set of large databases using massive parallelization on computer clusters.

In Chapter 2, the background part, the basic methodological (machine learning) and scientific (astronomical applications) concepts are explained. The respective sections are easy to understand for experienced readers of the respective fields. It is the connection between the two fields that comprises the main work of this thesis. The part on machine learning is very diverse and covers a wide range of tasks and concepts of machine learning algorithms. The astronomical introduction links the performed methodological work with current state-of-the-art astronomy and highlights the additional benefit gained by the use of a data-driven methodology. This thesis is written as a cumulative one and consequently, three publications are subsequently included as they were published in the respective journals. As usual for scientific publications, each of them contains an introduction and a methodological section which is just a compressed version of what is written in the background part.

The publications are presented in Chapter 3. In the first one, an alternative way of estimating the redshift of spectra by using the full information content available in the spectra is described. Therefore, the spectra are represented by a high-dimensional feature vector which solely contains information about the spectral features. Subsequently, distances between the spectra are calculated and eventually redshifts are inferred from the nearest neighbors. This allows a more general estimation of redshifts, as effects originating from the continuum can be efficiently suppressed.

The visualization of regularly sampled time series (brightness as a function of time) is the topic of the second publication. The central part of this work is to find a new two-dimensional representation of light curves that still contains the maximum information content. To that end, a new visualization algorithm is developed which can measure the prediction power of the compressed representation directly on the original data. The new algorithm is then employed on a number of regularly sampled and classified light-curves of an X-ray emitting binary black hole and a striking difference compared to classical visualization methods can be seen.

Most photometric survey are, however, ground-based and as a consequence, environmental effects prohibit a homogeneous sampling of the time sequences. The classification of irregular time series was, so far, done by representing the observations by a set of statistical features. With the presented work, the arbitrariness of the selection (and preprocessing) of the features can be avoided by describing the time series data as a probability density function. The new representation is very useful for transferring knowledge between databases as it is able to describe the photometric uncertainties correctly. Eventually, the presented methodology performs up-to-par with state-of-the-art feature-based classification methods and can thus be seen as a more general alternative to feature selection which can furthermore be used for visualization.

After the publications, Chapter 4 offers a discussion of the obtained results. There, the broader astronomical context and use of the newly introduced methodology is discussed in more detail. Finally, the results of this thesis are summarized in Chapter 5.

Chapter 2

Background

In this section, an extensive introduction to the science cases treated in the attached publications is given. The performed tasks are set into an astronomical context and prior work in the respective fields is discussed. Subsequently, an introduction to the topic of machine learning is given and the used concepts are explained in great detail.

2.1 Astronomical question to be solved with machine learning

The interplay between astronomy and machine learning is a win-win-situation for both sides. While astronomy provides an easy access to large and complex databases to the machine learning community, new analysis methods will increase the understanding of data inherent properties and allow a better access to the underlying physics.

2.1.1 Identifying outliers in spectroscopic databases

The detection of peculiarities in spectra can lead astronomy in a less biased way towards the discovery of new classes of objects and physics (e.g., Decarli *et al.*, 2010; Meusinger *et al.*, 2012). The focus of this work is on spectra which cannot be explained by a single-redshift system. This is a good tracer for outliers, per definitionem, only rarely represented in the SDSS database and can reveal undetected and unexpected behavior.

Sources of relative redshifts

In this section, the merging of supermassive black holes and the origin of gravitational lenses are described. While the origin of both processes are very different, they both are rare events and exhibit a relative shift of (one or several) spectral features with respect to the reference redshift. However, the relative shift between the two redshifted systems is severely different and can be used to discriminate the two different origins.

Supermassive black hole binaries (SMBHB) The interaction and merging of galaxies (see e.g., Toomre and Toomre, 1972) is a process which has been studied on hundreds of objects, with Messier 51 being probably the most famous one. The evolution and formation of galaxies in the universe is tightly correlated with merging events (White and Rees, 1978) which is also, more recently, supported by numerical simulations (Kannan *et al.*, 2015). In addition, Richstone *et al.* (1998) showed that (at least) the most massive galaxies host a super massive black hole (SMBH) in their center. Hence, the merging of the galactic centers and the SMBHs should be inevitable (Begelman *et al.*, 1980). Due to the conservation of angular momentum, the two (or more) merging SMBH will encounter a binary stage. The average distance of the

binary is continuously decreased by different effects on different scales that involve dynamical friction by gas, slingshot ejection of stars and eventually emission of gravitational waves (see e.g., Milosavljević and Merritt, 2003; Gualandris and Merritt, 2008; Popović, 2012, for more detailed descriptions of those processes). It remains unclear, what the typical time scales for the merging processes are, but as long as they are not dramatically short or long, numerous SMBHs in the binary stage should be detected. So far, only a handful of objects are known to be SMBHB and a few dozens are considered to be bona-fide candidates (e.g., Komossa *et al.*, 2003; Maness *et al.*, 2004; Valtonen *et al.*, 2008; Rodriguez *et al.*, 2009).

The detection and determination of the frequency of merging events can deliver valuable insights in the nature of galaxy formation and evolution. Additionally, the merging of SMBH may actually also explain the growth rate of the SMBH in the center of galaxies, which cannot be explained solely by accretion rates according to the Eddington limit. As studied in Sanders *et al.* (1988), the merging might also play a considerable role in the activation of galactic nuclei.

As the SMBH themselves do not emit any radiation, their detection has been limited to their interaction with the environment. In the Milky Way, the existence of a massive central region was inferred by studying the behavior of orbiting stars and clusters in its vicinity (Gillessen *et al.*, 2009). Another strong indication of the existence of SMBH are active galactic nuclei (see e.g., Antonucci, 1993). The observed extremely luminous emission originates from a very confined region and is most likely caused by accretion of surrounding material onto the SMBH. The emission is a composite of (multiple) black body components superposed by discrete narrow and broad atomic emission lines, such as the Balmer series or iron lines. In radio-loud galaxies the observed emission might be heavily influenced by jet activity which can cause a Doppler-boosting of the emitted light (Urry and Padovani, 1995).

To detect SMBHBs in large scale surveys, two distinct approaches have been used so far. The first one assumes that if the SMBHs are in binary stage, the entire system might exhibit brightness variations in a periodic fashion which can, for example, occur if one of the SMBH crosses the accretion disk of the other and thereby enhances its activity. The most famous example of a SMBHB detected by this method is OJ287 (Valtonen *et al.*, 2008), but also in a recent publication, (Graham *et al.*, 2015) SMBHB candidates have been found with this approach. Alternatively, one can assume that at least one of the SMBHs is active and due to the binary motion a shift between the two SMBHs or the SMBH and the host galaxy should be observed. This method has revealed another dozen good SMBHB candidates, but the selected samples are potentially also heavily contaminated by objects in which gas kinematics causes the double-peaked emission (Fu *et al.*, 2012). The latter approach will be discussed in more detail in a later subsection. It is, however, worth noting that the detection of any of the two mentioned effects does *not* proof the existence of a SMBHB. Further measurements, such as the change of the spectral shift over time (e.g., Liu *et al.*, 2014) need to be studied in order to confirm the existence of a SMBHB.

Gravitational lenses As predicted by the theory of general relativity, the line-of-sight alignment of two astronomical sources can, depending on the geometry of the alignment, cause gravitational lensing. The principle behind gravitational lensing is that the intermediate object is heavy enough to bend its surrounding space such that the deflection (lensing) of light can be measured. The consequences of gravitational lensing are manifold. Depending on the mass distribution of the lens, either multiple images of the source or arc-like rings can be observed. Some examples of gravitational lenses are shown in Figure 2.1. Gravitational lenses are a very important laboratory to address many kinds of astronomical questions. They are, for example, suitable scales to measure the absolute mass and the mass distribution of the lens (Kochanek, 1991). Additionally, they allow detailed insights on the lensed source as the spatial extent and

intensity of the source are magnified (Marshall *et al.*, 2007). Furthermore, by measuring the light travel times of different lensed source images (Kochanek, 2002) allows a direct estimate of the Hubble constant without the gauging effect of the distance ladder.

(a) *J1000+022* **(b)** *Abell 383* **(c)** *J1038+4849*

Figure 2.1: *Examples of gravitationally lensed objects. All lenses have been observed with the Hubble space teletscope. Credit: https://www.spacetelescope.org/*

If the lensed source images has a separation lower than the fiber diameter of a given spectrograph ($3''$ in case of SDSS), the acquired spectrum should contain contributions of the source and the lens. Since the lens is in-between the observer and the source, two different redshift system (one for the source, one for the lens) should be apparent in the spectrum as long as the brightness of the lens and of the lensed source image are of a comparable order. As the two sources are not gravitationally bound and the lensing effect only occurs for certain source-lens and lens-observer distances, it is expected that the two redshifts should differ quite significantly. Therefore, the difference in redshift can exceed unity. Consequently, the different redshift systems are easily separable in terms of spectral resolution, but it might be on the other hand more complicated to assign the correct redshifts to the respective systems. A huge advantage in the search for gravitational lenses is that they have been detected before (see e.g., Muñoz *et al.*, 1998; Bolton *et al.*, 2006) and thus the search strategies are more elaborate than for the SMBHB. Additionally, high-resolution imaging enables astronomers to validate possible candidates by resolving the separate components or detect arc-like structures.

Detection of multiple redshift systems

In Section 2.3, it will be shown that the SDSS database does not allow to extract these objects directly from the database. Only objects properties which obey expected behavior (in terms of templates) are reliable. This is the reason why several efforts have been made to detect peculiar objects in the raw spectra. Smith *et al.* (2010) focus on a relative shifts of the [OIII]-doublet. Bolton *et al.* (2006) analyze the residuals after removing the best-fitting spectral template and identify potential spectral lines. A more advanced and data-driven approach to detect shifts between H_β and $[OIII]$ lines has been presented by Tsalmantza *et al.* (2011). This approach seems also the most promising in terms of completeness and reliability, as it takes into account the measurement uncertainty and computes the likelihood of a double-redshift system. However, all of the approaches used so far, still depend on a model for the continuum and the lines. These (cf. Section 2.3) were inferred by performing a principal component analysis (PCA) on spectra that have been manually shifted to their rest frames. This can cause a significant drawback on the methodology, as the subtraction of the (wrong) continuum may create artifacts which might mimic features. Despite Bolton *et al.* (2006), all approaches focus on specific spectral features which imposes strong limitations on the flexibility of the approach. Furthermore, the approaches

in the literature have solely focused on shifts between emission lines, while absorption lines are indicative of the redshift of the galaxy as well and have been entirely ignored so far.

2.1.2 Analysis of time series data

With the automation of survey telescopes and the improvements in instrumentation and data processing, the number of photometric surveys increased dramatically[3] The observed variability of sources can have manifold origins. In the Galactic plane, the majority of variable sources are stars on the instability strip and eclipsing binaries. Outside the galactic plane supernovae or the variability of star-forming and active galaxies and quasi-stellar objects are the cause of variability. For both stellar and extragalactic origins, the time scale of variability ranges from hours (see e.g., Lindgren *et al.* (1975) for Wolf-Rayet stars and Heidt and Wagner (1996) for BL Lacs) to several years (see e.g., Soszyñski *et al.* (2009) for long-periodic variables and Valtonen *et al.* (2008) for the super massive binary black hole candidate OJ287).

As variability can be caused by a variety of physical effects, an analysis of a monochromatic light curve can only give hints about the origin of the variability. Graham *et al.* (2014) show, e.g., that light curves can indicate quite reliably whether an observed variability is of stellar origin or caused by a quasar. A detailed classification, however, requires more information like photometric colors or spectra. The imprints of the origin of the variability in stellar light curves are usually very clear and unique and therefore allow a more detailed classification. In Figure 2.2, light curves of two pulsating variables and three different types of eclipsing binaries are shown. All of those variable stars are periodic and thus, the visualization can be compressed from the original light curve (left side) to a phase-folded one. Given the correct period P for a light curve with photometric observations measured at observation times t, a light curve can be phase-folded by converting the time-axis to a phase

$$\phi = t/P - \text{floor}(t/P) \tag{2.1}$$

where the function floor(x) rounds x down to the nearest integer.

Variable stars

As aforementioned, this work focuses on variable stars. As mentioned in Watson (2006), stellar variability is generally divided into extrinsic and intrinsic variability. Extrinsic variability is caused by alignment effects (binaries, rotational variables, microlensing), intrinsic variability denotes that a changing physical state of the star is causing the variable behavior (pulsations, eruptions, bursts). In this part, the focus will be on pulsating and eclipsing variables.

Instability strip The instability strip (Gautschy and Saio, 1995, and references herein) denotes a region in the Hertzsprung-Russell-Diagram (HRD, Russell, 1914) where the pulsating variable stars are located. In Figure 2.3, the absolute magnitude, using the distance modulus and the parallax estimated by Hipparcos (Perryman *et al.*, 1997), is shown as a function of the color $B-V$. Due to the lack of a temperature estimate this plot reflects a mixture between the observable Color-Magnitude-Diagram (CMD) and an HRD. The gray dots in the background are a 2-dimensional density estimate of all stars observed by Hipparcos. The colored points are variable stars of different classes as extracted from General Catalog of Variable Stars (GCVS, Samus *et al.*, 2004). The location of the instability strip is marked by the black dashed lines[4]. The deviations from the instability strip, main sequence and the giant branch can be mainly attributed to the variability, which shifts objects vertically in the plot.

[3] A selection of existing open-access optical time series databases will be given in Table 2.1.

[4] It should be noted, that binaries are of course not pulsating variables, but are just plotted the same way.

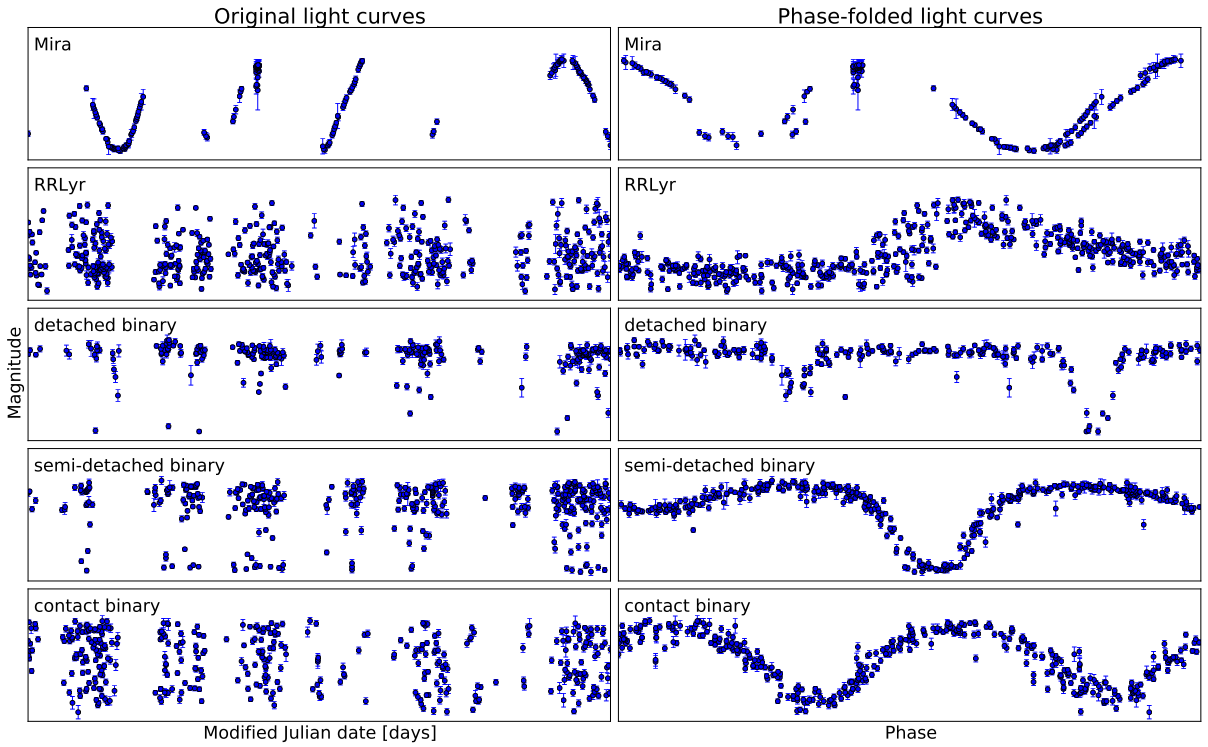


Figure 2.2: *Raw (left) and phase-folded (right) light curves for different types of stellar variability, denoted in the subplots, from the ASAS survey (Pojmanski, 1997).*

The location in the HRD and the shape of the phase-folded signals are the defining properties of the different variability classes. The underlying physical mechanism driving the pulsations is the same for all the classes in the instability strip. The photosphere of those stars contains Helium in different ionization stages located at different depths of the photosphere. Depending on the temperature of this location Helium is predominantly neutral (HeI), ionized once (HeII) or ionized twice (HeIII). When the stellar envelope starts contracting, the inner layers of the photosphere are heated. The fraction of HeIII in the HeII layer increases, which in turn causes the opacity of this layer to increase. Since now the radiation from the inner parts cannot escape the shell anymore, it is driven outwards due to the radiation pressure. As a consequence of the increasing distance, the temperature of the former HeII layer drops. Thus, the newly formed HeIII starts to recombine to HeII again. Since now the radiation pressure in this layer decreases strongly, the gravitational force attracts the outer shell again. This leads to a compression of the envelope which was the starting point of the pulsation. Even though the underlying mechanism behind the pulsation is the same for all stars in the instability strip, the different subclasses can vary quite strongly in period and amplitude.

The interest in pulsating stars is very high, since many different effects can be tested on them. The most prominent correlation is the period-luminosity dependence (Benedict *et al.*, 2007 for cepheids and Cáceres and Catelan, 2008 for RR Lyrae). This makes pulsating stars good standard candles in the local universe for distances that are too large to be measured with parallaxes but too close to see the effects of the cosmological expansion. In addition, the new discipline of astroseismology (Gilliland *et al.*, 2010) analyzes different pulsation modes and can thereby obtain detailed information on the interior of those pulsating stars. As a consequence, many stellar parameters, such as the radius, age and metallicity can be estimated very accurately.

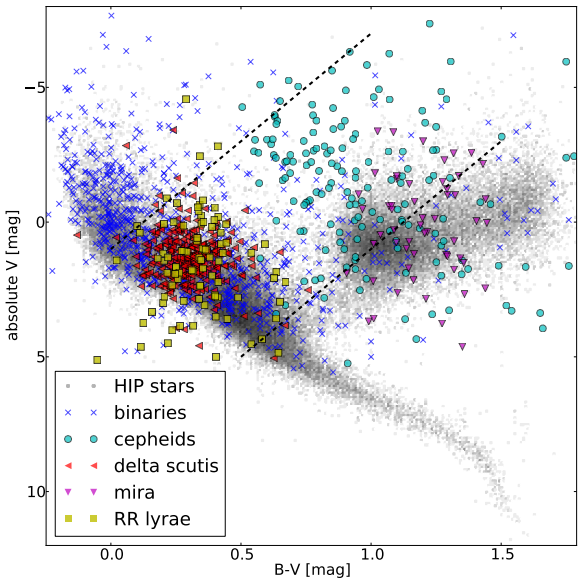


Figure 2.3: *Absolute magnitude (using Hipparcos parallaxes) versus color. The different types of variable stars (extracted from the General Catalog of Variable Stars, Samus* et al.*, 2004) are marked; the approximate location of the instability strip is highlighted by the black dashed lines.*

Binary stars The variability in binary systems originates from the mutual occultation of two stars or a star and a planet. Since the detection of exoplanets requires very high photometric precision and are easier detectable in regular observations they will not be considered any further. The shape of the variability of an eclipsing binary depends on multiple aspects, such as the radius and the spectral type of each of the stars. Additionally, the orbital parameters have a severe impact on the observed light curve, e.g. inclination.

From a physical perspective, a binary is in one of the following states: detached (ED), semi-detached (ESD) or contact binaries (EC). Those states are defined according to the Roche lobe of each star, which defines whether the tidal force from the other star exceeds the gravitational force of the object itself, and thus leads to an overflow of material. For semi-detached binaries one star fills its own Roche lobe and the other does not, for detached (contact) binaries none (both) fill their Roche lobe. Apart from the orbital and geometrical parameters also physical effects (e.g., limb darkening) influence the shape of the variability. For each of the classes, examples were shown in Figure 2.2.

A very special kind of binary stars are the X-ray binaries. In those binary systems a main-sequence or giant branch star orbits a compact object like a white dwarf, a neutron star or even a stellar black hole. These binary systems are discriminated according to the mass of the non-compact object (high/low) as then the material overflow is caused by different physical processes (stellar wind/Roche lobe overflow). Depending on the type of the compact object, the X-ray emission is caused by thermal emission of the accretion disk, formed by the conservation of angular momentum. An additional component originates if the compact object has a strong magnetic field. Then the ionized material from the accretion disk is channelled along the magnetic field lines and eventually hits the surface (so this does not happen for black holes) of the compact object with extremely high speed. This causes a heating on the surface (hot spot) which increases the local temperature to several million Kelvin. This causes black body radiation in the X-ray wavelengths. Consequently, the hardness of the X-ray radiation (frac-

tion of high-energetic to low-energetic X-ray emission) is strongly dependent on the type of the compact object.

The detection and analysis of binary stars is of great importance, since with given average distance and orbital period of the system, the summed absolute mass of the two binary stars can be determined. The frequency with which binaries occur constrains also the theory of stellar (and therefore also galactic) evolution. Additionally, the X-ray binaries provide good experiments for testing and validating accretion disk theories (Shakura and Sunyaev, 1973), which play also an important role in other fields of astronomy, e.g., for quasars.

2.1.3 Inspection of regularly sampled time data

The launch of two recent space observatories, COROT and Kepler (Auvergne *et al.*, 2009; Borucki *et al.*, 2010), have opened the window into a new era of time series data. Apart from Hipparcos (with a rather high cadence of 10 days), these are the first surveys to observe huge areas in the optical with a reasonably short cadence in regular time intervals. Those regularly sampled time series data offer two very important insights into stellar and extragalactic astronomy: they allow a long-term study of variability on short and long time scales and it enables scientists to treat time series data as a regular sequence. From a methodologically point of view, these can be handled in a more straight-forward way. The short cadence and high photometric precision of the observations over a long time interval are especially interesting for studying RR Lyrae stars or for occultations by small (earth-like) exoplanets which was the primary goal of both missions. While the selection of both fields in Kepler and Corot favor, following the mission goal, the observations of galactic stars, the regular observations permit a view not biased by observation strategy. Thus, the observed sources can be seen as a statistical ensemble which are "drawn" unbiased and thus should be relatively representative of the population of variable stars in our Galaxy. Consequently, the understanding and classification of the different sources in the observed field of view can have considerable impact on our understanding of the evolution of our own Galaxy.

For the analysis of the acquired data, new methodology needs to be developed as the sequential nature of the data bears huge advantages over the irregularly sampled ground-based light curves. Additionally, the photometric error of the space-based observations is significantly lower (due to the missing atmosphere) and thus better insights into the physical processes of the variable source can be gained. While the observations provide unique opportunities in understanding the variable nature of a huge variety of different variable sources, the analysis of the data volume produced by them is computationally demanding. For example, Kepler observed within its first complete public release Q1 (33.5 days) over 150,000 variable sources with a cadence of 29.4 min and thus producing over 240 million photometric measurements. It is evident, that this amount of data cannot be inspected manually and that definitely a pre-sorting has to be done. In principle, two basic ways exist to perform this pre-clustering; transferring knowledge from existing, ground-based time series database or cluster on the acquired data solely.

The transfer of knowledge seems on first sight to be easier applicable, as already available knowledge (experience) has to be only transferred. In reality, however, it is quite hard to perform this task. The knowledge to be transferred is either the classes of objects which have been observed in Kepler and another survey (label transfer), or the properties that have lead to this classification (knowledge-transfer). The label transfer is rather straight-forward to apply. However, the common sources between Kepler and, for example, the All Sky Automated Survey (ASAS) hardly exceed 1,000. Consequently, they do not provide a sufficient database for predicting these classes on Kepler data. The method of knowledge-transfer is, however, much harder to apply, as the observational biases of the different surveys as well as the very different photometric uncertainties have to be taken into account accordingly. It is therefore unclear how

the prediction quality of the class assignment decreases by predicting from a very irregularly to a regularly sampled survey with much shorter cadence.

Unsupervised methodology (more details are given in Section 2.4) can potentially be a good alternative, as biases from other classifiers and surveys are absent. Additionally, this could provide also a new understanding of the different existing classes of variability and should be in principle even able to highlight outliers. The application of unsupervised algorithms on time series data/sequences bears, however, other difficulties. In a first step, the raw sequences should be transformed to a vector representation, which should be insensitive to the median brightness of the compared objects and insensitive to shifts along the temporal axis. Due to the absence of labels for the Kepler database[5], the analysis of X-ray time series, with already assigned physical states, is preferred. For this, the X-ray binary system GRS1915+105 (a known microquasar) was monitored with the Rossi X-ray Timing Explorer (RXTE) by Greiner *et al.* (1996). Subsequently, Belloni *et al.* (2000) manually inspected all the observations and assigned to each time series one of 12 distinct classes. These are believed to be also different physical states[6] the binary system was observed in. This dataset provides the perfect testbed for the presented visualization algorithm.

2.1.4 Classification of irregular sampled light curves

The assignment of classes to a given set of irregularly sampled photometric data points is an extremely challenging task since many different parameters influence the temporal behavior of the source. The creation of models for different variability classes is very difficult for sparsely sampled light curves as they depend on (partially) degenerate parameters. The most important parameter for modeling (and also classifying) light curves is the period of the signal and huge efforts have been made to find a methodology which is suitable for all types of variability (see e.g., Stellingwerf, 1978; Scargle, 1982; Schwarzenberg-Czerny, 1989). In Graham *et al.* (2013) it was, however, shown that the dependency between type of variability and quality of the period estimate could not be broken by any of the proposed algorithms. Additionally, even the modeling of correctly phase-folded light curves with correctly assigned classes is difficult since even little deviations from the true underlying parameters cause huge deviations due to the repetitive structure of the underlying signal.

Classical way The classification of time series data was in the past treated in two different ways. The first implied prior knowledge about the observed source, e.g., from spectral or multi-band photometric data. Those acquired data were then modeled and the position in the HRD was thereby confined. This allowed to infer the different types of variability. In the second approach multiple domain experts were asked to label (phase-folded) light curves manually. The manual classification strongly relies on a correct determination of the period. It is also evident that the technical feasibility and limited human resources are severe drawbacks of these methods. It should be noted that if new variability classes occur later on, they remain undiscovered and are labeled wrongly.

First data-driven approaches To circumvent the problem of re-observing sources or engaging a large number of experts, alternative approaches have been investigated. Feeding extracted

[5]Labels have been provided by Debosscher *et al.* (2011). However, in the third publication it will be argued why these results should be handled with care. The main concern is that the very different sampling between the learning and testing set was not taken into account accordingly. Thus, the given labels should be rather seen as a hint but are far away from being reliable.

[6]Examples for two of the states are given in the respective publication.

feature vectors to classification algorithms (see, e.g., Hastie *et al.*, 2009) has gained much attention in recent works (Debosscher *et al.*, 2007; Richards *et al.*, 2011; Donalek *et al.*, 2013). The main idea has been to reduce the irregular light curves to homogeneous measures (called *features* hereafter) which are independent of the sampling and the phase of the light curve. Typical choices for the features are the period, the intensity of the respective periodic signal, median absolute deviation and many more (see e.g., Richards *et al.*, 2011, for a complete list). Those features are then used to train a classifier (see Section 2.4) that is able to assign labels to unlabeled data.

Weaknesses One of the big shortcomings of classical and learning-based approaches is, that the photometric errors are not taken into account accordingly. Neither the domain experts, nor the feature-based approach respects the (sometimes quite severe) uncertainty introduced by noisy measurements, the strong observational bias which is introduced by seasonal effects, and potential systematic bias of the observation strategies. It is interesting to notice that the number of observations correlates quite drastically with the labels. Since the methodology is not defined *generic* enough, it is not suited to transfer obtained knowledge between datasets directly. Transfer means that the classifier is trained on one database and is used to predict labels on another. Even though, this direct approach has been applied (Debosscher *et al.*, 2007), it should be noted that a lot of information which exists in the tested data is simply discarded, because the errors are not considered accordingly. The existing methodology also lacks a natural extension to regularly sampled time series. There the use of features is highly questionable as the sequential nature of the data is not considered.

2.2 Complex Data in Astronomical Databases

The amount of publicly available astronomical data has increased dramatically over the past decade. A major part of this data flood can be attributed to a rather manageable number of sky surveys, e.g., FIRST (Becker *et al.*, 1995), 2MASS (Skrutskie *et al.*, 2006) or SDSS (see, e.g., Ahn *et al.*, 2014). The concept behind those surveys is to observe and/or monitor fields or all of the sky in an automated and unbiased fashion. The scientific aims of the different surveys are very distinct and therefore, also the observation strategy and methodology (imaging, spectroscopy) differ significantly. In total, all these surveys produce a huge and rich variety of different data types over different wavelengths and (spatial and spectral) resolution. This variety is a blessing and a curse at the same time as it offers huge possibilities for answering scientific questions, but makes the application of analysis techniques computationally and methodologically more challenging. In order to understand the usability of a methodology on a given database, the databases have to be described by common properties, explained in the following.

2.2.1 Database characteristics

The characterization of a database is an inevitable step before a developed methodology can be applied. A common description of databases in big data science is given by the four V's: Volume (number of entities, dimensionality), Variety (homogeneity, dimensionality), Veracity (uncertainty, complexity) and Velocity. In the presented work, velocity is not an issue as only static databases are considered. With the launch of GAIA (Perryman *et al.*, 2001) or first light of LSST (Ivezi *et al.*, 2011), speed will become a major issue. All other properties are described in the following.

Homogeneity

A database is usually composed of N entities, each of which has a number of entries (features) in a database. If this number of entries has the same meaning and returns a valid number for all entities, the database is called *homogeneous*. An example of a homogeneous database is the photometric database of SDSS, as each object observed in the r band has corresponding values in the $ugiz$ bands as well. If an object is observed in a band, but not detected, it will be assigned a magnitude which is beyond the brightness limit of the SDSS. However, a non-detection is a valid entry as well. The homogeneity of a database is an important requirement for the application of data-driven approaches since only then the reliability of an approach can be judged in a uniform way throughout the sample.

In order to apply data-driven approaches to an inhomogeneous database, it has to be *homogenized*. A straight-forward example of homogenization is interpolation (e.g., by using physical models). For example, each spectrum in the spectral database of SDSS has, by its nature, a slightly offset dispersion solution, i.e., the first pixel does not always correspond to the exact same wavelength value. Since the offset from a general grid is fairly low, the entire spectrum can be interpolated (flux-conserving) to a more general grid. Time series databases, obtained by ground-based observations, are an example for fully inhomogeneous databases as the temporal sampling is subject to many environmental effects.

Dimensionality

The dimensionality of a database denotes how many *independent* entries (features) exist per entity. Usually, no prior knowledge is imposed on the database, all quantities (and even their combinations like, e.g., colors) are seen as independent properties. This assumption can obviously also prohibit a detailed understanding of processes if, e.g., each temporal measurement

of a time series is seen as an independent measurement even though two adjacent photometric measurements can be strongly correlated.

Uncertainty

The level of uncertainty in a database denotes how heavily an observed signal is truncated by noise. A common notion of the level of uncertainty in astronomy is the signal-to-noise ratio (SNR) where a $\mathrm{SNR} = 1$ indicates that the observed signal is of the same order as the statistical uncertainty of the signal. In astronomy, the uncertainty of a measurement depends on multiple factors that are independent. The introduced uncertainties can be categorized in three groups: physical, technical and analytical uncertainty. The physical uncertainty, only occurring in ground-based observations, originates from the time-dependent behavior of the atmosphere. The technical uncertainties are caused by the non-ideal technical infrastructure of the telescopes, e.g., misalignment, efficiencies and pixelization. Eventually, analytical uncertainties are either of a statistical nature (Poisson noise) or are artifacts that are introduced in the adjacent analysis of observations. The different uncertainties are indistinguishably superimposed and consequently, measurements performed under different circumstances (weather, hardware, brightness) can be compared directly only if the respective uncertainty is estimated and taken into account correctly. In feature extraction approaches, the inclusion of the uncertainty poses one of the major issues and will be discussed later on.

Complexity

The complexity of data refers to the underlying physical behavior of an observed source. An observed single-band image of a distant star can principally be considered a non-complex signal since it can be described with an analytical model (e.g., Moffat profile) and only the scaling has to be adapted. However, it should be noted that also the description of a point source can be very complex if the superimposed uncertainties are of a larger magnitude than the signal of the point source, such as for faint stars. An image of an extended nearby spiral galaxy is more complex as it is a superposition of an underlying behavior (e.g., Sersic profile) with an imposed structure (e.g., spiral arms) and many visible local sub-structures (e.g., clusters or supernovae), so that the overall model has many adjustable parameters.

2.2.2 Examples of astronomical databases

To understand the practical implications of the database characterizations, two different astronomical database types, which are also used in the publications, are presented.

Spectroscopic data

Spectroscopic data are complex, (nearly) homogeneous, noisy and high-dimensional data. Occasionally, parts of the spectra are missing due to technical problems, however, it is assumed that those can be interpolated linearly. An illustrative example of the different effects impairing the original (complex) physical measurements are shown in Figure 2.4. In this work, the Sloan Digital Sky Survey (SDSS, Ahn *et al.*, 2014) spectroscopic database is used, since it is one of the largest existing spectroscopic database at optical wavelengths. As the understanding of the data acquisition and processing of each of the delivered data is a key point in applying data-driven learning methods, the SDSS and its data reduction pipeline (that is to remove detector and night sky behavior) will be explained in more detail.

For the spectroscopic observations, a 2.5 m telescope located at Apache Point observatory is dedicated to the SDSS. The spectra are acquired by attaching optical fibers to pre-manufactured

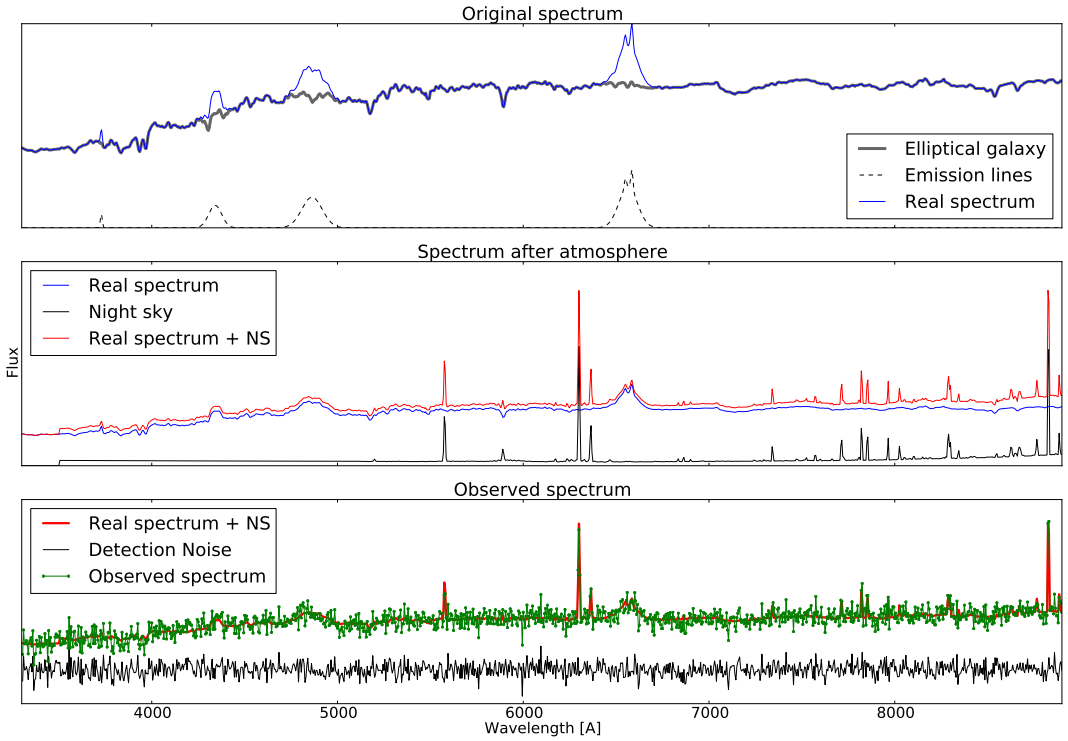


Figure 2.4: *Artificial optical spectrum (top) with impacts by night sky (center) and detector noise and discretization (bottom).*

plates which are designed for certain sky regions. Since the light is then transported via optical fibers (3″ in diameter) to the spectrograph, all spatial information is lost. The older spectrograph (SDSS) was able to record 640 spectra in parallel, the newly installed BOSS spectrograph can deal with over 1,000 spectra at a time. The reduction of the 2-dimensional spectral data into 1-dimensional spectra is described in detail in Stoughton *et al.* (2002). Here, only a short summary is given. The analysis of 1-dimensional spectra is described in Section 2.3.

The first step in the reduction of the spectra is the subtraction of the detector bias and the division by flat fields in order to correct pixel-to-pixel sensitivity variations as well as distortions along the optical path. In order to account for different hardware introduced effects, such as grating/mirror efficiency, standard stars are observed occasionally. It is assumed that the hardware behaves in a well defined manner between the observations and thus, all these effects are considered to be static and will not be considered any further. The most important and frequently changing impact on the spectra is the night sky emission that has to be accounted for. Extra fibers on each respective plate point to empty regions in order to estimate the contribution of the night sky. The SDSS pipeline then interpolates the night sky behavior to the positions of the observed sources. While this seems like a straight-forward task it happens occasionally, according to our publication in 1 out of 1,000 cases, that the night sky is heavily over-estimated. While the origin of this effect remains unclear, it shows that the data preprocessing can have a considerable impact on individual scientific questions, e.g., the identification of outliers.

Time series data

Time series (also called light curves) are repetitive photometric observations of distinct sky regions or of all the sky. Due to seasonal, atmospheric, and weather effects in ground-based (GB) observations, objects cannot be observed in a regular fashion. Consequently, GB light curves are

extremely inhomogeneous. The inhomogeneity does in this case not only refer to the irregular sampling of the observations but also to very different photometric errors, according to the observing conditions at the time (and location). As the average time between two measurements is typically much larger than the time scales of the variability of most of the objects, it is hardly possible to interpolate the data, as in the case of spectra. On the other hand, several space missions have been dedicated to the search for exoplanets and solar-like pulsational behavior. Therefore, regions in the Milky Way have been observed continuously with different cadences, delivering homogeneous time series data. Those datasets are extremely valuable for understanding the mechanisms of variable behavior and yield interesting insights in the biases introduced by the irregular observations performed by GB measurements. Table 2.1 gives an overview of some of the publicly available time series databases.

Name	Start	Bands	Area	Reference
Hipparcos^{S}	1989	optical	MW	Perryman *et al.* (1997)
MAssive Compact Halo Objects (MACHO)	1992	V,R	MW, LMC	Cook *et al.* (1995)
Optical Gravitational Lensing Experiment (OGLE)	1992	V,I	MW, SMC, LMC	Udalski *et al.* (1992)
All Sky Automated Survey (ASAS)	1997	V,I	all sky	Pojmanski (1997)
Robotic Optical Transient Search Experiment (ROTSE)	1998	unfiltered	all sky	Marshall *et al.* (1997)
Sloan Digital Sky Survey (SDSS)	1998	u,g,r,i,z	Stripe 82	Ivezić *et al.* (2007)
Catalina Real-Time Survey (CRTS)	2003	V	all sky	Drake *et al.* (2009)
COnvection, ROtation and planetary Transits (COROT)^{S}	2007	unfiltered	MW (field)	Auvergne *et al.* (2009)
Kepler mission^{S}	2009	unfiltered	MW (field)	Borucki *et al.* (2010)

Table 2.1: *A selection of existing optical photometric surveys with some selected attributes, like observed region (MW: Milky Way, MC: Magellanic Cloud) and filterbands. The three surveys observed with space telescopes are marked with an* $\boldsymbol{S}$.

2.3 Schemes of data analysis

As a consequence of the immense incoming data stream, the online analysis of data has become as important as the technical feasibility of the surveys themselves. Apart from the data reduction, the software needs to be able to analyze the obtained data in an extremely reliable fashion, as systematic deviations will affect all obtained results. The correctness of the analysis is validated only rarely, such that even small problems can have severe impacts on huge databases and remain potentially undetected.

The data analysis is usually science case dependent. However, most of the analysis techniques in astronomy belong to one of the following tasks

1. Clustering/Visualization (unsupervised)
2. Classification (supervised)
3. Regression (supervised)

where the tasks in brackets refer to the nomenclature in computer science whether at least parts of the data are labeled (supervised) or not (unsupervised). In the following, those tasks will be explained using the example of the SDSS spectral pipeline. This is described in further detail in Stoughton *et al.* (2002); SubbaRao *et al.* (2002).

2.3.1 Unsupervised approaches

In the absence of assigned labels for the database entities, clustering usually is the first step to be done in order to obtain an overview over the broad variety of the acquired data. Clustering denotes the detection of overdensities in the high-dimensional feature space and many approaches have been described in the literature to achieve this goal such as K-means (Lloyd, 1982) or DBSCAN (Ester *et al.*, 1996).

The simplest approach to get an impression of the different data present in SDSS could be to plot color-color diagrams over all observed objects. This type of visualization can already impart a good idea about existing structures within the data. Throughout this work, only visualization will be considered[7]. The goal of visualization is to project a given data set into a lower-dimensional space with the aim to visually inspect and cluster those data according to existing substructures. While for the five bands *ugriz* this is an achievable task (even though, those are for all color-color and color-magnitude diagrams already more than 100 plots), higher dimensional data require a more efficient dimensionality reduction that allows an inspection in lower-dimensional space. An important aspect of this dimensionality reduction is that distances from the high-dimensional space should be preserved as well as possible in the lower-dimensional projection. Visualization is, however, not part of the SDSS pipeline. Instead, clustering was performed in a very rudimentary, and therefore also very subjective, way by handing a set of 2,200 spectra to domain experts with the task to *cluster* them manually (Vanden Berk *et al.*, 2001). The classes obtained from the different experts are then merged into common classes and to each of them a template is assigned. In total, 33 unique classes defined by a respective template spectra were obtained. All those different classes are considered to be independent classes such that for a given target spectrum only one class can be assigned. The majority of those templates describe stars on the main sequence or in later evolutionary states (e.g., white dwarfs) nearly all of which could be described by a more principle model composed of a black body radiation with superimposed spectral lines. It is therefore questionable how well the chosen templates reflect the true behavior of the different classes (stars, galaxies, quasars). A strong

[7] In this work, visualization and dimensionality reduction are used interchangeably.

imbalance between the frequency of those classes in the data (25.7%, 63.7%, 10.6%) compared to the fraction of templates attributed to them (69.7%, 18.2%, 12.1%) is apparent.

2.3.2 Classification

Classification denotes the assignment of a discrete class (in this case a template) to an unlabeled spectrum. For this purpose, a set of training objects with known labels is required. With the help of those, labels for unseen data can be assigned using different approaches, such as support vector machines (Chang, 2011) or random forests (Breiman, 2001).

As shown in Figure 2.4, SDSS spectra are a superposition of an underlying continuum and superimposed spectral absorption/emission features. To get an idea of the importance of the different spectral features, all spectra that have been used for clustering were shifted with respect to their manually assigned redshift into their rest frames. In a next step, a principal component analysis was performed on those in order to obtain the most important spectral features apparent in the learning set. A list of those features can be found in the Appendix A.1.

In order to assign classes to unlabeled spectra, the templates were then combined with the principle components (which are freely variable in amplitude, but constrained in broadness) in a way such that the deviation between the model and data points is minimized (least-square fitting). In order to account for the shifting due to the redshift of a source, the optimization of the redshift is performed on a grid, simultaneously. The template with the lowest deviation was then assigned to be the class of the tested spectrum.

2.3.3 Regression

Regression denotes the estimation of a continuous property, such as redshift or slope. While classification denotes the assignment of discrete classes, regression is usually performed on physical parameters (such as redshift) but in principle the same methodology as for classification can be used.

In case of the SDSS spectra, the simultaneous fitting of template, redshift and principle components to an unseen spectrum, delivers numerous regression values. Apart from redshifts, line widths, amplitudes, many more parameters are extracted and stored in a database. This database is accessible via a standard *Structured Query Language* (*SQL*) interface under http://dr10.sdss3.org/. Thus, the highly complex spectra are converted into a catalog of simple descriptors. This is a very handy way of describing spectra which are well represented by any of the given templates. However, it remains questionable how helpful such a database structure is when it comes to real scientific questions.

2.3.4 Shortcomings

The above considerations make it evident that there exists a degeneracy between the class template, redshift and amplitude of the main components. This implies that if the fitting procedure fails in any of the tasks, all of the respective properties and classes will be assigned wrongly. In order to identify these misfitted spectra one can study the behavior of the residuals; the "SMALL_DELTA_CHI2" flag indicates whether several solutions with comparable reduced chi-squared values exist. This flag is activated for more than 15% of all analyzed spectra! Consequently, a manual investigation of all the misfitted spectra is impossible. In Figure 2.5, an illustrative example is given that shows that the value of residuals is not a sufficient tracer for truly suspect spectra.

The organization of the database is well suited for the detection and analysis of behavior that can be well described by the chosen set of templates together with the superimposed main components. However, the provided database does not allow access to information which is not

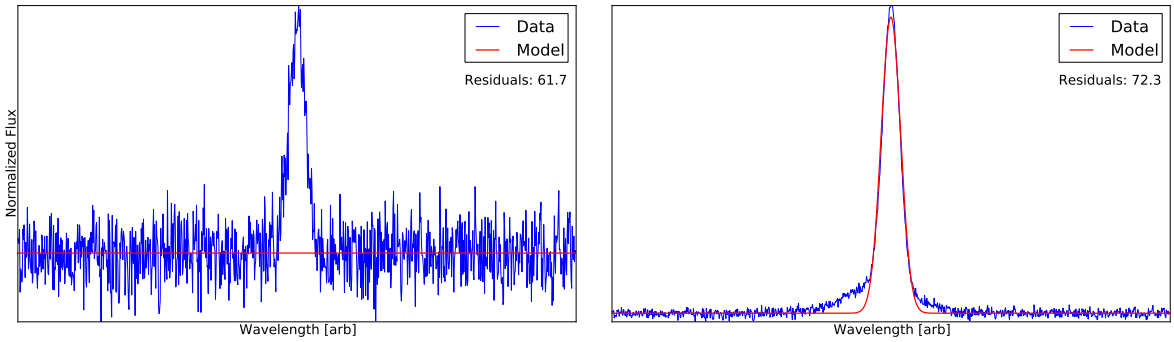


Figure 2.5: *Artificial data highlighting drawbacks of solely minimizing the residuals. On the left side the feature was entirely ignored by the fit, but due to the lower signal-to-noise ratio, the impact on the residual is lower than for the fairly well prescription of the more prominent feature on the right side.*

encoded in the applied model. This lead many astronomers to a reinspection of the raw data in order to select spectra which fulfill a given requirement. For example, Collinge *et al.* (2005) were aiming to identify feature-free power-law spectra in the SDSS spectral database. As no catalog feature existed, that allowed an explicit query, the entire database was mined just for this task. In a differently organized database, the authors could have pointed out some objects, which they identified to be interesting candidates and the database could have returned a set of spectra which obey similar properties. The selection based on similarity instead of extracted features is also a more natural way in terms of human perception.

Eventually, it is important to understand why the application of data-driven methodology could be advantageous over *static* models. In the following, models will be discriminated into static and dynamic (learning) models. Static models do not adapt to the data provided and thus, remain fixed throughout the analysis. On the contrary, dynamic models take into account the data they have been provided and adapt to them accordingly. The SDSS reduction pipeline is an example of a static model. This choice entails three important consequences:

1. The underlying (static) model is not reproducible.

   As the expert team decided on a common set of templates using their expertise, it is impossible for a neutral person to come up with a set of templates which is similar to the presented one. Given a fixed dataset, a data-driven approach could actually learn a set of templates which is reproducible by anybody.

2. The selection of the (static) templates cannot be validated.

   The choice of templates is solely subject to the expertise of the investigators. Only one set of templates has been provided and thus, it cannot be validated whether a better/more general set of templates could exist. If, however, the model is generated by the data, the given dataset can be split into different subsets from which the model(s) can learn. Subsequently, the best model is chosen on the basis of an objective function[8] .

[8]That is the criterion judging the quality of a model, a commonly used measure is the root-mean-square of the residuals.

3. The (static) model will, independent of the incoming data, remain unchanged.

   In case a given (potentially unknown) class was underrepresented in the dataset handed to the experts, it will never be represented in form of templates, even if the occurrence of the class is extremely frequent in the newly incoming data. For the static models, the only way out is a manual interaction, i.e., introduction of a new template. The dynamic model will learn the behavior and weight the importance of the newly arisen subset according to its frequency.

All the points discussed in this section will be explained in greater detail in the following Section 2.4. This is an easy-to-read introduction to the mentioned dynamic (or data-driven) models.

2.4 Learning from data

Machine learning is a modern statistical approach for hypothesis testing by creating, validating and selecting and rejecting models. This strategy was pursued automatically by humans in the past. Experiments were performed and the acquired data lead to a fundamental understanding of processes occurring in Nature, e.g., Newton's laws. With the advance of new technology, experiments and, a new theoretical background, some of this knowledge has been revised and/or adjusted such that, e.g., Newton's laws are now seen as a special case of Einstein's general relativity theory. The understanding of science has been either driven by experiments that contradicted existing theories or by predictions made from theories and later-on (not) confirmed by experiments. With the advances in computer technology, each of those paradigms (experimental and theoretical) was supplemented with a computer-driven extension, data-driven science and simulations, respectively. Machine learning learning offers a new approach to validate existing models in huge database allows to develop new models, according to the supplied data. Additionally, the detection of outliers has always been of great interest in astronomy, as they can challenge the integrity of the underlying models.

In order to highlight the differences in the process of model selection performed visually and data-driven, a simple two-dimensional clustering task is considered in Figure 2.6. The respective datapoints are drawn from three Gaussian distributions whose centers are highlighted in the last frame. With each step (from left to right) more and more datapoints are drawn from these distributions. At each step a dashed line separating the potentially existing clusters in the given data is drawn. The number and course of these separating lines is subject to the choice of the individual investigator and infinite different lines could be drawn that would separate the hypothetical clusters equivalently well. One can immediately see that the human decision making is very flexible, e.g., a new class can be introduced easily and old decisions can be revised. The price for this flexibility is that the choice of the clusters is irreproducible, as anyone might see different classes in the respective subplots. Additionally, visual inspection can be performed

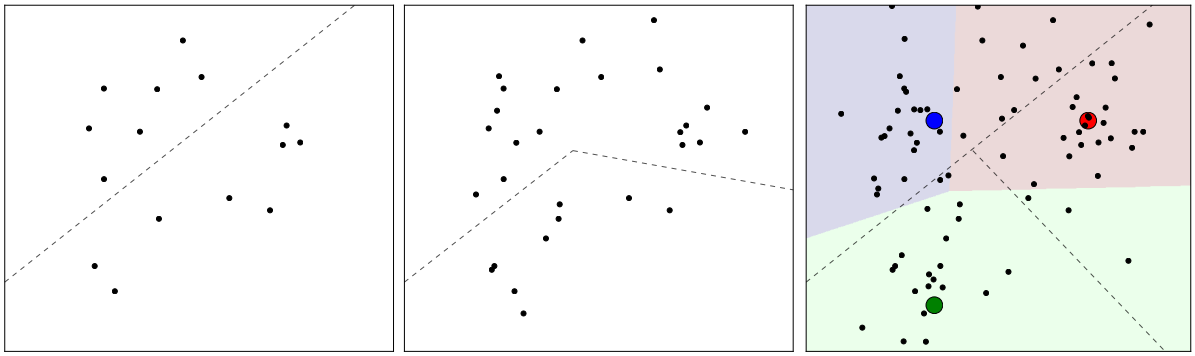


Figure 2.6: *The principle of visual adaptive clustering. From the left to the right more random points are drawn from three Gaussians, whose centers are indicated in the last plot with large colored dots. The human-based clustering is very flexible and can adapt easily to newly occurring clusters. However, the visual clustering is very subjective and is thus not easily reproducible. For comparison, a clustering based on a basic data-driven clustering algorithm (K-means, MacQueen, 1967) with three centers is shown in the last frame as background color.*

only due to the sufficiently low dimensionality of the data. Probably in three, definitely in four dimensions more than one plot have to be created in order to obtain a fundamental understanding of the relations between the different dimensions. In high-dimensional data, a visual inspection is impossible to perform and alternative approaches need to be developed that are efficient and

better reproducible than the human inspection. The training of flexible algorithms (adaptive learning) is state-of-the-art research (Zliobaite *et al.*, 2012), but will not be investigated in this thesis. Instead the application of existing machine learning methods on astronomical data is focused. The use of data-driven methodology requires the definition of an objective criterion that evaluates the quality of a given model. Different definitions of this objective criterion exist and obviously its choice has a severe impact on the performance of the learning algorithm. This objective criterion should reflect what similarity means in the context of the respective science case. A more detailed description of the concepts of model selection and data similarity will be given subsequently.

In the following subsections, some fundamental problems arising from the use of deterministic and purely model-based approaches are highlighted. As a consequence, new methodology as provided by statistical learning is introduced and the advantages and disadvantages of those are discussed. The structure and content of this section are mainly inspired by Bishop (2006); Ivezić *et al.* (2014). It is the aim of this short introduction to outline some of the concepts of machine learning and probability theory.

2.4.1 Model selection

Model selection is the process of judging the quality of each of a set of candidate models on a provided dataset. Two central definitions are of great importance for model selection: model complexity and validation. To highlight the meaning of those terms a typical regression problem is considered. The data (green dots in Figure 2.7) are generated from the sigmoidal logistic function

$$\frac{1}{1+e^{-x}} \tag{2.2}$$

by drawing samples of x from a uniform distribution between $[0, 1)$. Gaussian noise with a mean $\mu = 0$ and a standard deviation of $\sigma = 0.1$ is added. The task is now to fit a function f which interpolates between the data points. The fit is optimized by minimizing a given objective function, e.g., the widely used least square measure:

$$E(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}) = \sum_{i=1}^{n} \left(f\left(\boldsymbol{w}, x\right) - y_i\right)^2 = \sum_{i=1}^{n} \left(\left(\sum_{j=1}^{\theta} w_j \mathcal{N}\left(x_i - \mu_j, \sigma\right) \right) - y_i \right)^2 \tag{2.3}$$

where $f\left(\boldsymbol{w}, x\right)$ is a continuous function which is linear in the weights $\boldsymbol{w} \in \mathbb{R}^{\theta}$. In the presented case, $f\left(\boldsymbol{w}, x\right)$ is a mixture of Gaussians with θ^{th} degree and the width σ is adapted (but fixed for all j) according to θ. This minimization will obviously favor high values of θ since, as soon as $\theta \gg n$, the objective will eventually reduce to zero. This fact is called over-fitting. The over-fitting occurs since the number of degrees of freedom of the model exceeds the number of data points and consequently, the model can imitate any abrupt behavior. However, most correlations between physical quantities are smooth rather than abrupt. Therefore, a low fitting error does not necessarily imply a good prediction quality. In order to measure the prediction quality, the sample of data points has to be split into a training and a validation set. Here, the k-fold cross-validation is used, i.e., the dataset is split into k equally-sized folds. The model is then optimized (trained) on $k-1$ folds in order to predict on the data from the k^{th} fold. The cross-validation measures the generalization of a given model and, thus, the abruptness of a model will effectively be penalized as the performance is evaluated on unseen (test) data instead of optimizing the model only on training data. In Figure 2.7, the difference between the least square error (deviation from green points) and the validation error (deviation from blue points) are given. The former one favors the highest degrees in the function f, while the validation error,

which is a measure of the prediction quality, indicates that a degree of $\theta = 4$ is sufficient/ideal to describe the underlying shape.

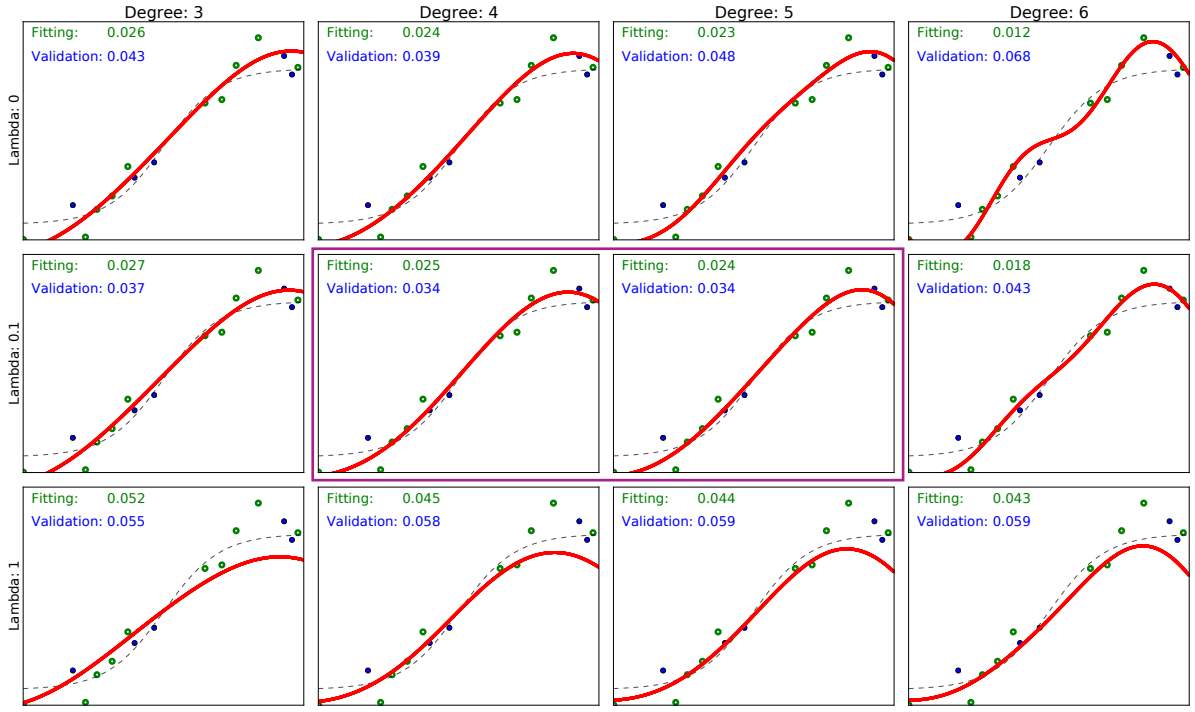


Figure 2.7: *Randomly drawn training (green) and validation (blue) points from a sigmoid function (gray dashed line) with normal noise. The fit (red line) is performed with a mixture of Gaussians with θ components (increasing in horizontal axis) and with increasing regularization values λ towards the bottom. The two best predicting models are highlighted by a purple box.*

Another way of limiting the model complexity θ is the introduction of a regularization term λ such that the error function becomes

$$E_{\text{regularized}}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}, \theta) \equiv E(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{w}, \theta) + \lambda \boldsymbol{w}^T \boldsymbol{w}. \tag{2.4}$$

The insertion of the parameter λ leads to a penalty on large weights and thus, smoother models are strongly favored over abrupt ones. The price for the penalization is that another free parameter λ arises which has to be optimized for each problem, individually. In Figure 2.7, it can also be seen that an intermediate regularization (middle row) yields a lower validation error and delivers a more suitable model. On the other hand, a high value of λ can lead to over-simplification (lowest row), while a too low choice does not constrain the complexity of the model at all. The combination of validation and regularization can limit the complexity of the given model. As a consequence, the smoother model can more reliably predict formerly unseen points and therefore yields a better prediction performance.

2.4.2 Similarity and data representation

The notion of (dis-)similarity is the basis for all the following methods and is therefore discussed in more detail. A typical choice for the similarity between two *vectors* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ is

$$D(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} \tag{2.5}$$

which transforms into the Euclidean (Manhattan) distance for $p = 2$ ($p = 1$). Apart from the exponent in the definition, also the individual dimensions of the vector might be weighted to highlight or suppress the importance of a certain vector component. By definition, all of the dimensions of the vectors are treated thereby independently. Other similarity measures can also include the interaction between non-independent dimensions. However, instead of utilizing a notion of similarity, it is more desirable to obtain *distances* between vectors. A similarity measure is called a distance if it obeys the following conditions:

1. $D(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ and only 0, if $\boldsymbol{x} = \boldsymbol{y}$
2. $D(\boldsymbol{x}, \boldsymbol{y}) = D(\boldsymbol{y}, \boldsymbol{x})$
3. $D(\boldsymbol{x}, \boldsymbol{z}) \leq D(\boldsymbol{x}, \boldsymbol{y}) + D(\boldsymbol{y}, \boldsymbol{z})$.

The similarities defined so far are tailored for measuring distances between two vectors. Other measures of similarity can be used if two probability *densities* have to be compared. While Equation 2.5 also provides a valid distance for probability densities it does not take into account that the given densities are normalized. Therefore, tailored similarity measures between densities, such as the Kullback-Leibler divergence and the Bhattacharyya distance, exist. Throughout this work, similarity between vectors and densities will always be measured with an according distance. The choice of the distance has a considerable impact on the final results. The modification and adaption of the distance measure (e.g., by introducing dimensional weights in Equation 2.5) has a model character as well; this is called metric learning.

In this work, only distances between vectors and distributions are considered. In the subsequent publications, conversions of data that are not vectors or densities themselves into such, are investigated. This conversion bears, however, many risks as the intermediate model, converting non-vectorial to vectorial data, has to follow quite naturally from the presented data type.

The SDSS spectra, for which the redshift is to be estimated, are of a vectorial nature already[9]. However, to provide the regression approach only with information that is relevant for estimating the redshift, a preprocessing (e.g., subtraction of the continuum) is applied to the spectra. This step reflects more a data manipulation than a real conversion, however, it is inevitable in order to be able to estimate redshifts.

In the other two science cases, light curves which are of a sequential nature are considered. For visualizing regularly sampled, variable length and potentially time-shifted signals the sequences are converted into a fixed-length vector representation. This is done by employing a model to the sequences. This new representation then allows to compare the sequences. On the contrary, in the third publication the sequential behavior of the irregularly sampled light curves is discarded and the observed photometric datapoints are converted into a probability density. As mentioned earlier, several ways exist to measure distances between those densities.

2.4.3 Curse of dimensionality

The similarity measures introduced in the previous part have a well understood meaning in low dimensions. However, with increasing dimensionality of the data, the classical notion of distance looses its relevance. This is the *curse of dimensionality*. Data with an enhanced dimensionality require a more complex model to be described accordingly. This imposes strong limitations on the use of data- and model-driven approaches on these high-dimensional data as they have to account for very detailed behavior of the data.

[9]This is only true, because dependencies between different wavelengths have been neglected. In fact, the measurements between the pixels are correlated by the physical processes of the object.

To visualize the problem, a simple two class classification task is considered. Given that in a dataset each item is represented by a four-dimensional vector[10], the dimensionality of the data is four. To perform a classification, the high-dimensional space can be subdivided into regular bins and in each bin the majority of contained objects dictates the label of an formerly unseen object. Assuming that all quantities are normalized between $[0, 1)$, the four-dimensional space can be subdivided into $\left(\frac{1}{\Delta}\right)^d$ sub bins if a bin width of Δ is desired. For a width of $\Delta = 0.01$ those are already 100 million sub-bins! Thus, already four dimensions can be seen as a high-dimensional classification problem; in the presented work, the treatment of feature vectors with $100 - 1,000$ dimensions will be discussed. This leads to a computationally infeasible complexity. Apart from this, the classification in the geometrical subdivision method depends strongly on the chosen length and, even more severely, the high-dimensional space usually contains large areas not populated by data.

To avoid the geometrical separation of the high-dimensional space into a larger number of bins, one can either use a different measure for estimating the local density at a given point or one can project the initial high-dimensional vector into a lower-dimensional space. Examples for both approaches are the basis for the presented stduies and are explained in greater detail in the following.

2.4.4 Local density estimations

In order to explain the concept of local density estimations, the two-class classification problem, introduced in Section 2.4.3, is considered again. Instead of investigating the local neighborhood by dividing the high-dimensional feature space into hypercubes, the continuous definition of the density is reviewed. The probability of a value x occurring in a region $\mathcal{R}$ can be computed by

$$P\left(x \in \mathcal{R}\right) = \int_{\mathcal{R}} p\left(x\right) dx \tag{2.6}$$

where $p\left(x\right)$ is the probability distribution of x. For a large enough number of N observations, $K \approx NP$ data points will be contained in the region $\mathcal{R}$. If the region $\mathcal{R}$ is small enough, the probability distribution $p\left(x\right)$ can be considered to be locally flat and therefore constant[11]. Hence, Equation 2.6 can be approximated by $P \approx p\left(x\right) V$. Combining this, the probability distribution can be estimated by

$$p\left(x\right) = \frac{K}{NV} \tag{2.7}$$

with V being the volume enclosed by $\mathcal{R}$. To estimate the local density for a given dataset either the volume V or the number of investigated objects N can be fixed to a constant. This gives rise to support vector machines and the k-nearest neighbors approach, respectively.

Support vector machine

In the support vector machine (SVM), the volume to be investigated is defined by the radius R of the chosen kernel. The kernel assigns a high importance to objects that are close in (a given) distance compared to the kernel radius R. Typical choices for the kernel[12] in a SVM are the rational square or the radial basis function

$$RBF(x_i, x_j) = e^{-\frac{\left(x_i - x_j\right)^2}{R^2}} \tag{2.8}$$

[10] This corresponds in an astronomical meaning to 4 different measurements, e.g., colors.

[11] It should be noted at this point, that the assumption of a large enough volume and local flatness are actually *contradictory*, however, for an approximate estimation they can be satisfied simultaneously.

[12] Note that the use of the boxcar function will turn the SVM into a histogram with bin width R, however, it is discontinuous and thus not suited for mathematical optimization.

which will be used in the publication.

The SVM is a distance-based method projecting a given distance matrix into a higher dimensional space where the different classes are more easily separable. In this projected space, the SVM is trained by finding hyperplanes that separate the classes optimally. As soon as the projected dimension is chosen high enough, any dataset can be optimally separated. In order to avoid over-fitting in the low-dimensional space, a regularization term C may be introduced in the objective function. The error parameter C[13] is comparable to the inverse of the regularization parameter λ in Equation 2.4. It reflects the trade-off between high (low) model complexity and over-(under-) fitting of the training data for high (low) values of C. As in regularization, this is a parameter which has to be fine-tuned for a given classification problem by performing a grid search over C. Besides classification, the SVM can also be used for regression tasks.

k-nearest neighbors

Instead of fixing the volume to be investigated, it is also possible to estimate the probability density by choosing the number of data points k to be investigated. For this, distances (or dissimilarities) to the nearest datapoints are measured. From these, information about the tested data item is induced. A visualization of the two-class classification problem is shown in Figure 2.8. The concept of inspecting the local neighborhood can be easily extended to perform clustering/visualization (K-means) and regression tasks. In contrast to the SVM, the kNN is not necessarily convergent but comes with a higher classification accuracy.

2.4.5 Dimensionality reduction

Instead of dealing with high-dimensional data, a dimensionality reduction algorithm can be applied that reduces a feature vector to its most important (not necessarily meaningful) characteristics. Dimensionality reduction is a process which projects a high-dimensional vector into a lower dimensional representation. The performance of the dimensionality reduction can be measured by the loss of information caused by the reduction. Ideally, the algorithm would try to preserve the distances between two items in the high-dimensional space also in the lower dimensional representation. Due to the confinement in lower dimensions, the high-dimensional distances can only be preserved approximately, because the information content that can be stored in the lower-dimensional representation is limited.

There are multiple ways to perform a reduction of dimensionality and the used methodology strongly depends on the data to be reduced and the overall goal of the dimensionality reduction. The simplest data reduction algorithm (and most prominent one in astronomy) is to discard data. For example, Collinge *et al.* (2005) search for BL Lac candidates by mining the SDSS database for featureless spectra. Because also some stellar sources (DC white dwarfs) enter via this selection criterion, all sources are plotted in a color-color-diagram ($g-r$ versus $r-i$) and a separation is introduced by visual inspection. Effectively, the authors discarded all knowledge encoded in the other 8 color combinations and might, therefore, lose valuable information and, even worse, create a significant selection bias.

Another very frequently used reduction algorithm is the principal component analysis (PCA, Hotelling, 1933) where for a given dataset the main components contributing to the signal are extracted and presented as basis functions. The main drawback of this method is that it is a linear model and is not able to describe more complex behavior. Additionally, the PCA is based on the minimization of χ^2 and consequently, focuses mainly on the conservation of distances between dissimilar points. To preserve local distances, the t-distributed stochastic

[13] In this work, a ν-SVM will be used. The value ν is just a normalized version of C and therefore, the grid optimization of it can be performed between 0 and 1.

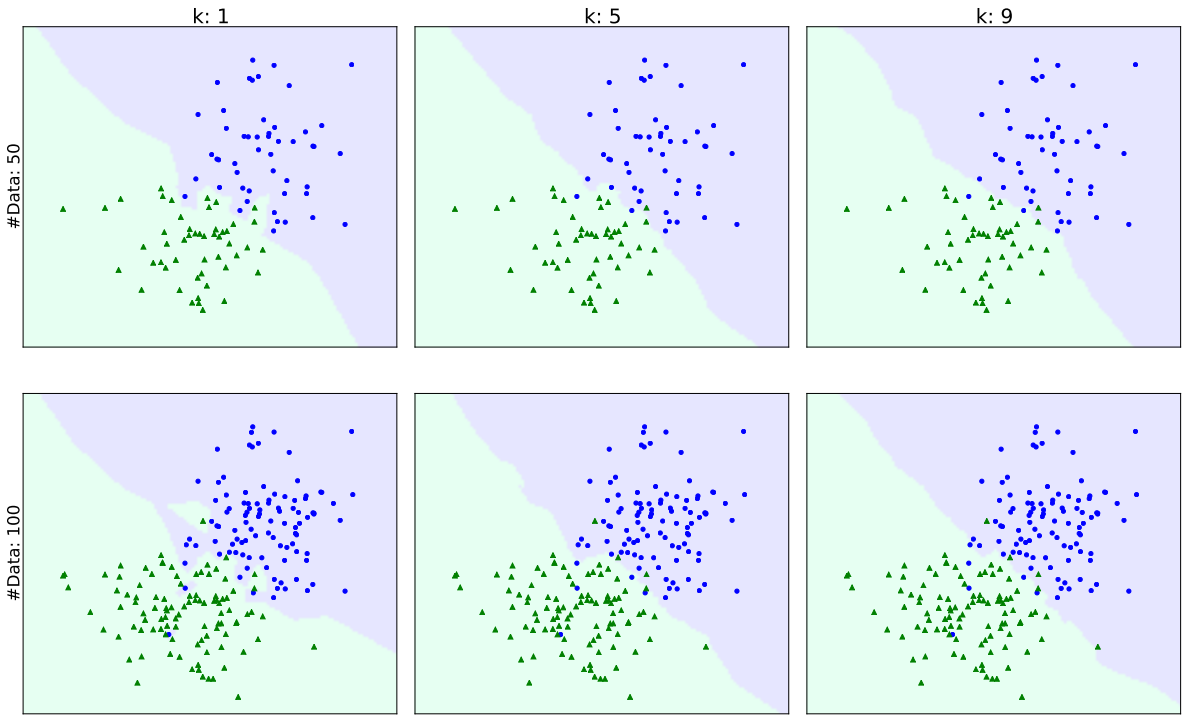


Figure 2.8: *Classification depending on different choices of k in a nearest neighbors classifier. The background color encoding highlights the course of the classification boundary, which is very abrupt (smooth) for low (high) values of k.*

neighbor embedding (abbreviated by t-SNE, van der Maaten and Hinton, 2008) can be used. The principal idea is to preserve the local density of each point by minimizing the Kullback-Leibler-divergence between the probability density in the high- and low-dimensional space.

Another dimensionality reduction algorithm is the autoencoder (Kramer, 1991) which is also used in this work. The autoencoder is a special artificial neural network (ANN) where the high-dimensional structure is represented by a low-dimensional hidden layer (bottleneck). In order to understand the working principle of the autoencoder, a fundamental understanding of ANN is required, which is provided in the following.

Artificial neural networks

The simplest form of an artificial neural network (ANN) is the 2-layer network (following nomenclature in Bishop, 2006), which consists of an input layer, a hidden layer, and an output layer. Each layer consists of a set of nodes, the number of nodes in the input and output layer is fixed by the dimension of the input vector and the desired dimensions of the output vector, respectively. For the connection between each two nodes a weight and an activation function are assigned. The activation function is usually represented by a step-like potential which causes a node to react or not, However, the optimization requires a continuous function which is then reflected by a non-linear, sigmoidal function, such as the sigmoid (Equation 2.2) or the tangens hyperbolicus. Thus, the ANN can allow for very complex and non-linear behavior which on one side is suitable for very complicated underlying models but also bears the risk of over-fitting.

The mathematical structure of the ANN is explained for the example of d-dimensional input

vectors (e.g., d=3 photometric measurements of 2MASS) and a 1-dimensional real-valued output vector (e.g., inferred redshift). A sketch of the proposed ANN is shown in Figure 2.9.

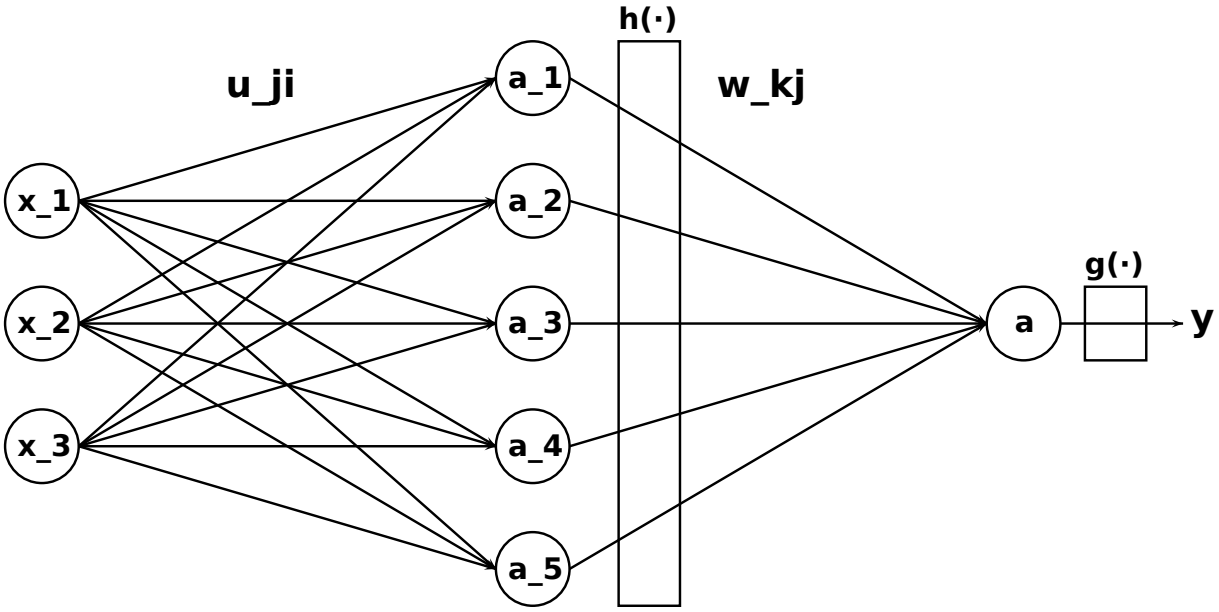


Figure 2.9: *Sketch of an artificial neural network with 3 input dimensions, 1 hidden layer with 5 neurons and a 1-dimensional output y_t.*

Each dimension of an input vector $\boldsymbol{x} \in \mathbb{R}^d$ is fed to its respective input node i. Subsequently, the activation value a_j for each node of the hidden layer is computed with

$$a_j = \sum_{i=1}^{d} u_{ji} x_i \tag{2.9}$$

where u_{ji} is the weight matrix for the hidden layer. The activation value is then processed through the hidden activation function $h(\cdot)$ to obtain the *hidden unit* $z_j = h(a_j)$. The output activations are then computed

$$a_k = \sum_{j=1}^{m} w_{kj} z_j \tag{2.10}$$

with the output weights w_{kj}, where m is the dimension of the hidden layer ($m = 5$ in the presented case). The regression value is then obtained by employing the output activation function $g(\cdot)$ to the output activations $y = y_k = g(a_k)$ where y has only one component in the example. Combining all the steps from before the presented regression problem can be computed as

$$y = g\left(\sum_{j=1}^{m} w_{kj} h\left(\sum_{i=1}^{d} u_{ji} x_i\right)\right). \tag{2.11}$$

The ANN is then trained by optimizing Equation 2.3 with respect to the input u_{ji} and output w_{kj} weights. Already in the presented example, 20 free parameters (weights) have to be optimized simultaneously. This optimization is very costly and in order to speed up the training, back-propagation is used which utilizes the difference between computed and expected output to update the weights. A more detailed description is given in Bishop (2006). As aforementioned, the ANN is a non-linear model with huge flexibility and thus, instead of optimizing solely Equation 2.3, a regularization as in Equation 2.4 can be introduced in order to avoid over-fitting.

Autoencoder

The working principle of the ANN was explained in the preceding section with the help of a regression problem. The concept of this is easier to understand because the training of the weights is equivalent to minimizing the objective function 2.4. The autoencoder is a very simple neural network which can be used for dimensionality reduction. It consists of a combination of an input, a hidden and an output layer. The dimensionality of the input layer is fixed by the dimensions of the provided input (say D), the number of neurons in the hidden layer is freely tunable (H), the number of the output neurons is dictated by the required output dimensionality, two in this case. To allow the autoencoder to measure the reconstruction error of the original input, the described structure is *mirrored* (cf. Figure 2.10). Thereby a bottleneck (blue) is created. The autoencoder is now trained by optimizing all intermediate weights such that the reconstruction error between the initial parameter representation θ and the reconstructed one $\tilde{\theta}$

$$\|\theta - \tilde{\theta}\|^2 \tag{2.12}$$

is minimized. The optimization of the autoencoder is quite time consuming as the algorithm has $2 \cdot ((D \cdot H)(H \cdot 2))$ free parameters. Additionally, again a least-square algorithm is used to optimize the autoencoder and, thus, more weight is put on preserving large distances instead of local ones, as for the PCA. On the other hand, the autoencoder allows a direct reconstruction (decoding) of the initial weights from a pair of given two-dimensional coordinates. This decoding part will be used in the presented visualization algorithm.

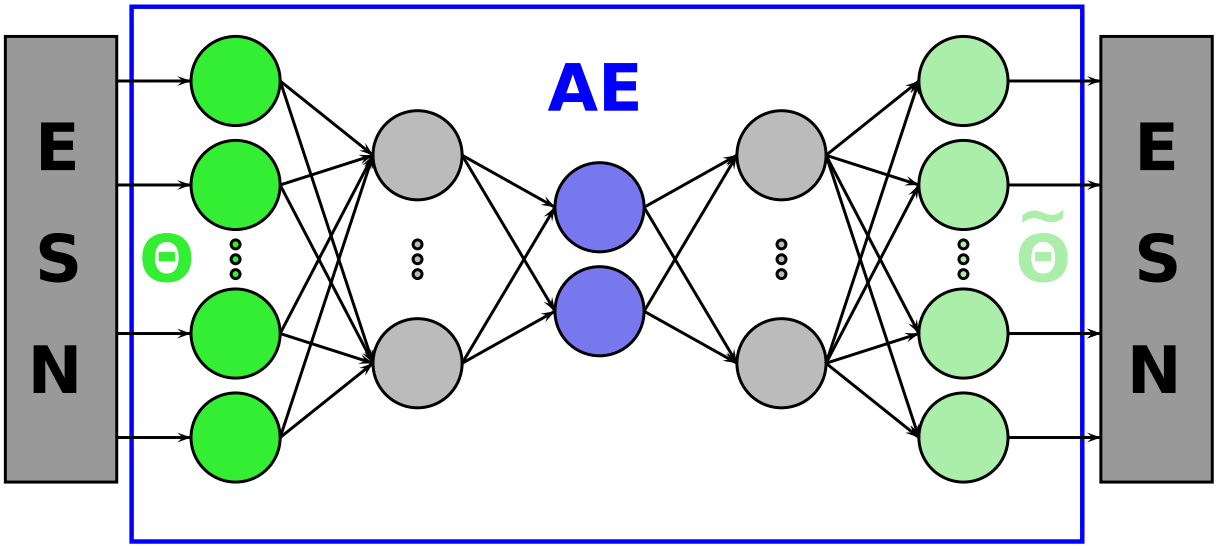


Figure 2.10: *Sketch of an autoencoder (AE), reducing the dimension of the output weights of the ESN (explained in the following) to two. The plain autoencoder would try to minimize the difference between θ and $\tilde{\theta}$. In the coupled version the reconstruction error on the data itself is measured instead by feeding the reconstructed weights $\tilde{\theta}$ to the same ESN.*

Recurrent neural network

As aforementioned, dimensionality reduction algorithms are solely meaningful when they are employed to vector data. To convert the provided regularly sampled light curves into vectors, a recurrent neural network is used.

The ANN presented in the former section was a fully-connected one. Different architectures of neural nets exist which include multiple hidden layers and different degrees of connectivity.

The focus of this section will, however, be on the recurrent network architecture. The concept behind a recurrent architecture is that it includes feedback connections in order to have access to computations of the previous time steps. An illustration of this can be found in Figure 2.11. The big advantage of the chosen architecture is, that it is able to trace the dynamical behavior of the observed system. This description is very suitable for approximating time series or sequential data (e.g., such as text). In the science case presented later on, it turned out that especially a very sparsely connected version of the recurrent network, the echo state network (ESN), is a good descriptor for the inspected time series. Another huge advantage of the ESN is that only the output weights w_{lk} have to be trained and are used for prediction. The input weights u_{ji} and the weights connecting the cyclic nodes v_{kj} are set to fixed values, which in turn define a reservoir. It is worth noting, that the best values for u, v can be found using a grid search over them for a representative training sample. Subsequent to this training, they are fixed throughout the visualization process. A big advantage of the ESN is that it is invariant against time shifts and returns for time series of variable length a fixed-length representation and is thus tailored for employing it on astronomical data.

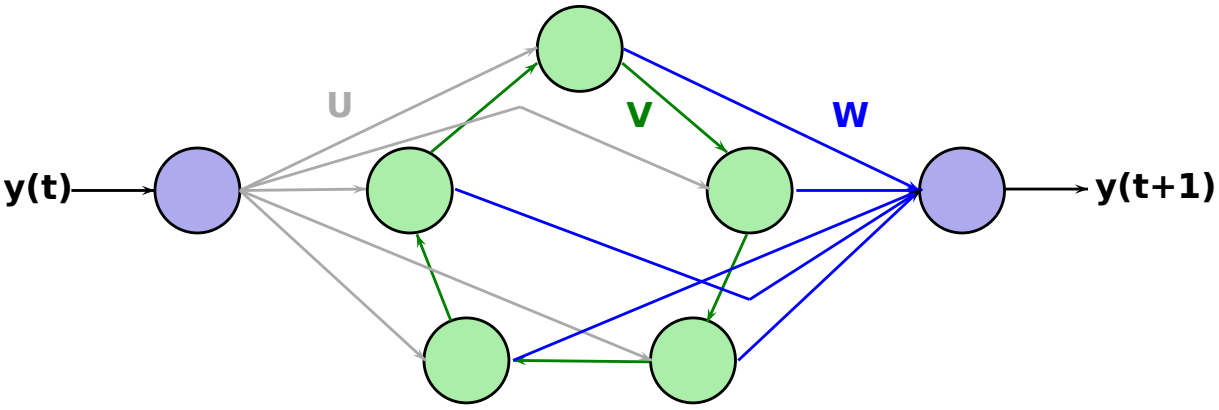


Figure 2.11: *Sketch of an echo state network with a hidden layer containing 5 nodes connected in a directed circle (red).*

Model-coupled visualization

The idea of the proposed visualization is now to use the ESN in combination with the autoencoder (AE). Instead of minimizing Equation 2.12, which measures the similarity between the original and reconstructed model parameters, a new objective is formulated. The idea is to measure the predicition quality of the reconstructed parameters on the raw data and thus the objective function is changing to

$$\|y - \tilde{y}\|^2 = \|y - f(\tilde{\theta})\|^2. \tag{2.13}$$

This newly defined objective function gives a much more intuitive feeling of similarity, as it does not measure the numerical similarity between the parameter vectors θ_A and θ_B of object A and B. Instead it measures how well a set of model parameters learned on A can also predict on B with respect to a given model. This circumvents additionally the problem of parameter renormalization which occurs when model parameters are compared, e.g., temperature and scaling of blackbody radiation are not of the same order of magnitude and would therefore be weighted differently.

While here only (regularly sampled) time series data are visualized, the concept can be extended to any (physical) model and visualization algorithm as long as they obey the following characteristics

1. Visualization algorithm has to be able to encode & decode
2. Model has to return a fixed-length vector representation of raw data
3. Model has to be able to make predictions on data given a fixed-length vector

While many models fulfill these requirements, the model should obviously be tailored for the supplied data in order to obtain a meaningful visualization. These models can be physically motivated (e.g., stellar spectra) or can be of a general kind (e.g., the proposed ESN for time series data). Additionally, the model should be non-trivial since, for example, a simple blackbody law can be easily represented with a single dimension (temperature). On the other hand, a complete stellar spectrum, including emission and absorption lines of different widths, has a considerable number of free parameters and is therefore tailored to be visualized. The number of visualization algorithms that allow decoding, that is the projection from the lower dimensional space into higher dimensional space, is strongly limited. Besides the used autoencoder, only the different versions of the PCA (like probabilistic, non-linear or extreme PCA) are potential alternatives for the visualization.

2.5 Contributions to the respective publications

The initial science driver behind the first study was the detection of new SMBHB candidates. I thought that the existing methodology implied some strong limitations on the SMBHB search. K. L. Polsterer provided general ideas to resolve this question and I investigated several of them. My contribution comprised the data acquisition, the adjustment of the data (preprocessing) and the methodology to the given problem. Subsequently, I implemented the concept in Python and started the parallelization of the program. The script was executed on a HADOOP-cluster, kindly provided by the third author M. Hoecker. Finally, I interpreted and discussed the obtained results in great detail and exclusively wrote the paper. There, I was also in charge of replying to the referee's concerns.

The basic idea of coupling a visualization algorithm to a fixed model and thereby reconstructing the prediction error was developed by the leading author N. Gianniotis. In order to highlight the importance of the presented algorithm, I helped to acquire datasets and to apply the new algorithm, by interpreting and discussing the scientific side of the experiment. P. Tiňo and K. L. Polsterer provided supervision over the work, R. Misra contributed the RTXE data. Additionally, I assisted with the implementation and improvement of the algorithm by experimenting with different visualization methods. The explicit inclusion of astronomical data into the experiments helps to highlight the special importance of this work to astronomy. For large astronomical surveys, such as COROT and Kepler, a generalized approach to investigate the nature of the huge variety of objects is urgently needed and under current investigation. Due to its importance in computer science, this publication was recommended for publication in "Neurocomputing".

My intention for starting the work described in the third publication was to avoid the arbitrariness of feature selection in light curve classification. I thought about generalizing the concepts of features and thereby developed the principle of static light curves which do not require features for their description. I acquired and preprocessed the data and implemented and parallelized the algorithm. N. Gianniotis contributed the theoretical fundament to this work by introducing me to different measures of density similarity. Again, K. L. Polsterer provided me with supervision and the respective computer architecture for the experiment. Subsequently, I acquired the data and discussed the results of the experiments. Finally, I wrote major parts of the publications and was responsible for handling the referee responses.

Chapter 3

Publications

This thesis is written in cumulative form. According to §7.2 of the PhD regulations in Astronomy & Physics (dated 10th of April, 2014) of the University of Heidelberg, three publications are required to submit a cumulative thesis. In the course of my PhD studies, I was engaged in the publication of the following articles which are the basis of this thesis and are included as they have been published in the respective journals. None of these publications was or will be used in a cumulative thesis of another co-author.

Publication I
Title: "Determining spectroscopic redshifts by using k nearest neighbor regression. I. Description of method and analysis"
Authors: Sven Dennis Kügler, Kai Lars Polsterer, Maximilian Hoecker
This article has been accepted for publication in Astronomy & Astrophysics (A&A)

Publication II
Title: "Autoencoding Time Series for Visualisation"
Authors: N. Gianniotis, D. Kügler, P. Tino, K. Polsterer, R. Misra
This article has been accepted for publication in ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.

The publication is accessible via www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2015-37.pdf

Publication III
Title: "Featureless Classification of Light Curves"
Authors: Sven Dennis Kügler, Nikolaos Gianniotis, Kai Lars Polsterer
This article has been accepted for publication in Monthly Notices of the Royal Astronomical Society (MNRAS)©: 2015, 451 (4), 3385-3392 Published by Oxford University Press on behalf of the Royal Astronomical Society.

A&A 576, A132 (2015)
DOI: 10.1051/0004-6361/201424801

Astronomy & Astrophysics

Determining spectroscopic redshifts by using *k* nearest neighbor regression

I. Description of method and analysis

S. D. Kügler, K. Polsterer, and M. Hoecker

Heidelberger Institut für Theoretische Studien (HITS), Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany
e-mail: dennis.kuegler@h-its.org

Received 13 August 2014 / Accepted 6 February 2015

ABSTRACT

Context. In astronomy, new approaches to process and analyze the exponentially increasing amount of data are inevitable. For spectra, such as in the Sloan Digital Sky Survey spectral database, usually templates of well-known classes are used for classification. In case the fitting of a template fails, wrong spectral properties (e.g. redshift) are derived. Validation of the derived properties is the key to understand the caveats of the template-based method.
Aims. In this paper we present a method for statistically computing the redshift z based on a similarity approach. This allows us to determine redshifts in spectra for emission and absorption features without using any predefined model. Additionally, we show how to determine the redshift based on single features. As a consequence we are, for example, able to filter objects that show multiple redshift components.
Methods. The redshift calculation is performed by comparing predefined regions in the spectra and individually applying a nearest neighbor regression model to each predefined emission and absorption region.
Results. The choice of the model parameters controls the quality and the completeness of the redshifts. For ≈90% of the analyzed 16 000 spectra of our reference and test sample, a certain redshift can be computed that is comparable to the completeness of SDSS (96%). The redshift calculation yields a precision for every individually tested feature that is comparable to the overall precision of the redshifts of SDSS. Using the new method to compute redshifts, we could also identify 14 spectra with a significant shift between emission and absorption or between emission and emission lines. The results already show the immense power of this simple machine-learning approach for investigating huge databases such as the SDSS.

Key words. methods: data analysis – astronomical databases: miscellaneous – methods: statistical – galaxies: distances and redshifts – catalogs

1. Introduction

In the past decades, the rapidly increasing amount of available data has been one of the greatest challenges in astronomy. In contrast to the amount of data, the number of techniques and the knowledge of how to analyze these large data sets has only increased slowly over time. When the first digital, photometric all-sky surveys were performed, the amount of available data was already too large to be inspected manually. With the advent of spectroscopic surveys and additional photometric surveys in multiple wavelengths, the available data volume increased so rapidly that novel approaches are mandatory.

So far, the most successful survey in astronomy has been the Sloan Digital Sky Survey (SDSS, York et al. 2000), which in its current 10th data release (DR10, Ahn et al. 2014) contains photometry for one billion objects and spectra covering the near-UV to the near-IR for roughly three million objects. In the future, surveys such as the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST, Cui et al. 2012) will reach this amount of data in a fraction of the time needed by SDSS. Thus more advanced techniques for handling those immense data streams have to be developed.

The determination of spectral redshifts and classifications of the SDSS spectra is based on template fitting. Therefore generalized templates are created by combining spectra of similar objects for all empirically determined classes of objects. By fitting those templates to the spectra, a number of predefined properties, such as redshift, can be individually computed for every object. The best fitting template is determined by applying all available templates to the data while allowing for some variation in a set of parameters (e.g. width of features) and testing the reliability of every model by computing a reduced χ^2. Instead of using the full information available, just a simplified model with a limited flexibility is applied, which does not allow a more detailed discussion of individual properties. Furthermore, the choice of the reference spectra and the creation of these templates have a strong impact on the determined properties.

With this publication we want to emphasize the power of statistical learning in huge spectral databases. From here on, "huge" refers to a large number of entities and dimensions. While this approach can in principle be applied to any database, we focus on SDSS. There are many applications of machine-learning techniques in astronomy (see Borne 2009; Ball & Brunner 2010). So far, spectroscopically derived properties have mainly

been used as ground truth to estimate redshifts on photometric data, for example in Laurino et al. (2011), Gieseke et al. (2011), Polsterer et al. (2013). In contrast less attention has been paid to the application of machine learning to the spectral data itself (see Richards et al. 2009; Meusinger et al. 2012), which can be mainly attributed to the "curse of dimensionality" (see Bellman & Bellman 1961). The ultimate goal would be to obtain spectral properties that are not based on the created templates but rather on the rich experience existing in the database instead.

The algorithm presented in this paper will perform a consistency check of the redshift calculated by the SDSS pipeline. We therefore assume that the majority of the spectra is fairly well described by one of the templates, so the redshift is determined to be reasonably precise. Of course the templates do not describe all kinds of objects perfectly, thus at least some will be misfit. The great improvement in calculating redshifts based on a data-driven approach is that the redshifts can be determined model-independent. This method is suitable for determining redshifts of unknown spectra and in a forth-coming paper we will present a value-added catalog of redshifts to the existing SDSS spectra. In this paper we focus on the technical side and explain the impact of the choice of different model parameters. To highlight the power of this new method, some outliers in terms of redshift in the used subsample are presented. The motivation for employing new methods for redshift computation is manifold:

1. *Validation*: cross-validating the self-consistency of the computed redshifts is crucial for understanding caveats of the SDSS pipeline. The independent determination of a redshift increases the confidence and the number of reliable redshifts.
2. *Calculating redshifts*: we are able to determine model-independent redshifts of existing and future spectra with high precision. This is possible since we are determining the redshift as an ensemble property and thus the theoretical resolution can be improved statistically with the number of similar spectra in the reference database, as well as with the dimension of the feature vector.
3. *Rare objects*: many different attempts have been performed to find rare objects showing shifts between spectral features in the SDSS spectral database (see Bolton et al. 2004; Tsalmantza et al. 2011). With the presented method we will be able to detect more of those since our method can deal with lower signal-to-noise ratio (S/N) than with template fits.
4. *Unexpected behavior*: this can be caused by objects of a previously unknown class or by a superposition of two classes. Those objects might possibly be the science drivers in the near future. Also, artifacts in the reduction pipeline/in the data can be discovered.

The paper is structured as follows. Section 2 describes the data used for creating and testing our model. In Sect. 3 we explain the basic approach used in our method in more detail. In Sect. 4 we discuss the performance of our method in terms of precision and reliability. Also some outliers and peculiar objects are discussed in more detail. A summary and an outlook follow in Sect. 5. In a follow-up paper, we describe the value-added catalog that gives redshifts for all available objects based on specific spectral regions. Additionally, a catalog containing all detected outliers will be presented there.

2. The SDSS spectroscopic database

For testing our method, we are analyzing the spectroscopic database of SDSS. This survey uses a dedicated 2.5 m mirror telescope located at the Apache Point Observatory (New Mexico, USA) to map the northern galactic cap and is a joint project by USA, Japan, Korea, and Germany.

The telescope was first used to image different stripes of the northern hemisphere in five filter bands using the drift scan method. Subsequently, interesting objects were selected by brightness limits and different color cuts for spectroscopy ($R = \lambda/\Delta\lambda \approx 2000$) with $3600\,\text{Å} \leq \lambda \leq 10\,000\,\text{Å}$ (Eisenstein et al. 2001; Richards et al. 2002; Strauss et al. 2002). Those selection criteria have a direct impact on the quality of reference sample. In the current DR10 (Ahn et al. 2014) more than three million spectra were taken of which far more than two million are nonstellar sources according to the SDSS-classification.

It is important to mention that depending on the applied learning technique, a large number of reference objects with a representative sampling is mandatory. With millions of objects, the SDSS is more than sufficiently large[1].

2.1. Data calibration/SDSS pipeline

As mentioned in the caveats of SDSS, the night sky subtraction can suffer from severe inaccuracy by rapidly changing conditions, e.g., auroral activity. Thus the night sky subtraction leaves a severe signature in some of the spectra, which is sometimes not taken into account correctly in the error estimation. As a consequence, faint features in the vicinity of strong night sky emission lines might be artifacts. The spectra are automatically labeled, both flux- and wavelength-calibrated, and eventually combined with potentially pre-existing observed spectra of the same object.

In a second step, the calibrated spectra were processed via an identification pipeline that assigned a redshift, a classification, and a velocity dispersion to the individual spectra (Bolton et al. 2012). The classification and redshift determination was performed with a principal component analysis (PCA) of a rest-frame shifted training sample. A linear combination of eigenspectra were then shifted with respect to flux and wavelength until a minimal residual was reached. The precision of the redshift for a single line is limited by the resolution per pixel ($\sim 100\,\text{km}\,\text{s}^{-1}$) of the spectrograph but can be improved by computing it independently for all lines that are available. This method is extremely efficient for spectra that show the expected behavior and as confirmed by performing a self-consistency check later on, and the quality of the SDSS redshifts has high reliability.

2.2. Reference and test sample

The analysis of the method was performed on a small subsample of the SDSS data in order to make the different model and parameter evaluations computationally feasible. The analysis of the algorithm is limited to the plates 0266 to 0289, including the exposures of all modified Julian dates (MJDs). Additionally, the sample was restricted to the redshift range between $0.01 \leq z \leq 0.5$. The selected restriction allows a more reliable prediction of the regression value because the density of reference targets in the direct neighborhood is sufficiently high. The chosen subsample includes 16 049 spectra in total. The redshift distribution of the spectra can be found in Fig. 1. In the following, this sample

[1] This statement is not only valid for the in-sample method presented here but also for the application on other datasets, as long as the wavelength coverage and the target selection criteria are comparable. This is because the data complexity of the reference sample, SDSS in this case, remains the same and thus a comparable number of references is needed for similiar precision.

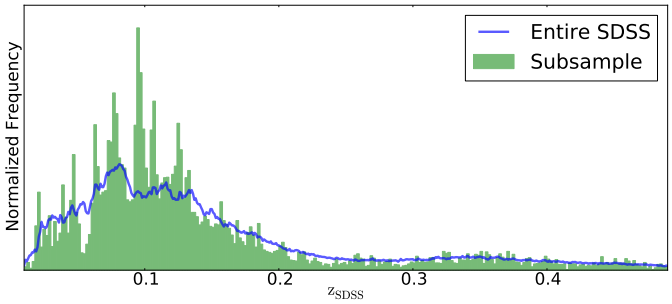


Fig. 1. Comparison of the redshift distribution of the selected subsample (green) and the entire SDSS (blue). There is a steep drop in the frequency toward redshifts $z > 0.25$. Single redshift bins are apparently undersampled.

is used as reference and test set at the same time; that is, we will perform a leave-one-out cross validation. That means that all but the target spectrum are reference spectra. Since we are only able to compute redshifts within the covered feature space, under-represented objects (high-redshifted galaxies, QSO) will yield worse redshifts than normally represented redshifts.

3. Applied method

The basic idea for determining the spectroscopic redshift z is to perform a comparison between similar objects. This is done by finding objects that look similar in terms of Euclidean distance and then computing the regression value of the unknown target by comparing it to the redshifts of the most similar spectra.

To be able to compare the spectra, instead of using the plain SDSS spectra we have to pre-process them. The method is a purely data-driven approach without deriving a generalization, and thus the quality of the redshifts relies directly on the chosen reference sample. While this seems contradictory on first sight, the method performs comparably on a smaller but representative reference set. It is obvious that the choice of a representative reference sample can only be obtained when domain-knowledge is included. Limiting the reference sample in redshift space would limit the derived values, respectively.

3.1. k nearest neighbor regression

Our method is based on k nearest neighbor (kNN) regression, which is a commonly used technique in statistical learning (Hastie et al. 2009). All spectra have d datapoints (corresponding to the individual flux measurements in the spectra) and are thus members of a d-dimensional feature space. The reference sample $\mathcal{R}$ consists of m entities and corresponds to the number of reference spectra from which the model learns, 16 048 in this case. Mathematically this sample can be described with

$$\mathcal{R} = ((\overline{x}_1, y_1), \ldots, (\overline{x}_m, y_m)) \in \mathbb{R}^d \times \mathbb{R} \tag{1}$$

where $\overline{x}_i$ is the ith d-dimensional input vector (spectrum under consideration) corresponding to the flux value in each pixel, and y_i is the redshift value z assigned by the SDSS pipeline.

The kNN regression is based on calculating similarities in the d-dimensional feature space. For any d-dimensional feature vector $\overline{s}$, the similarity to a reference object $\overline{x}_i$ can be estimated with distance measure $\Delta(\overline{x}_i, \overline{s})$. The most commonly used metrics are

$$\Delta(\overline{x}_i, \overline{s}) = \left(\sum_{j=1}^{d} |\overline{x}_{ij} - \overline{s}_j|^p \right)^{1/p} := \begin{cases} \text{Manhattan} & \text{for } p = 1 \\ \text{Euclidean} & \text{for } p = 2 \\ \text{Minkowski} & \text{otherwise.} \end{cases}$$

The impact of the choice of the metric on the final results was only marginal. Therefore we only use the common Euclidean distance. In general, the neighborhood $N_k(\overline{s})$ is determined on the basis of the representation of the reference objects $\overline{x}_i$ in the feature space, such that

$$y(\overline{s}) = \frac{1}{k} \sum_{\overline{x}_i \epsilon N_k(\overline{s})} y_i = \underset{\overline{x}_i \epsilon N_k(\overline{s})}{\text{mean}} (y_i) \tag{2}$$

however, here we make use of a modified version:

$$y(\overline{s}) = \underset{\overline{x}_i \epsilon N_k(\overline{s})}{\text{median}} (y_i) \,. \tag{3}$$

Different algorithms exist for finding the k most similar spectra $N_k(\overline{s})$. The most straight-forward one is the brute-force method where each spectrum is simply compared to each of the others and the distance is computed. In contrast, spatial structures exist (kd-, ball-trees) that are able to structure the data in advance. The average time to find the closest spectra is thus significantly lower once the search structure is created. When experimenting with spatial trees, we learned that the dimension of our data is apparently so high and the data themselves are so unstructured that spatial trees do not perform significantly better than the brute-force method, and as a consequence only the brute force method is used throughout the paper and for the future catalog.

The considered kNN regression is limited to interpolating values within the reference sample. As a consequence, redshifts of objects with extremely high redshift or very peculiar spectral features cannot be determined correctly.

3.2. Requirements

The method of kNN regression can only work efficiently if the following requirements are met:

1. The majority of the redshift determinations by SDSS is correct.

   In the following the deviation of the SDSS redshifts in comparison to the correct redshift is assumed to be small. This is verified by comparing our results to the redshifts determined by SDSS. One has to keep in mind that for a large fraction of the data, the template fitting works quite well, and the redshifts are fairly reliable.

2. The number of objects in the reference data set is large compared to the dimensionality.

   This is already met in our test subsample. Nonetheless this is quite surprising because the number of entities is approximately the number of dimensions (4000). It appears that the multidimensional feature space is sufficiently homogeneously populated with reference objects. Applying this method to the entire database will just strengthen that assumption further.

3. It is possible to distinguish noise from real signals for most of the data.

   This requirement is harder to meet because the distinction between signals and noise, especially for low S/N spectral lines, has always been a huge challenge for astronomers. In this work, we use an approach that is based on a simple similarity measure used by the type of the applied

regression method. The basic assumption is that when a detectable line exists anywhere in the spectrum, it should be possible to find similar spectra that, within their errors, have a similar redshift. Those form a sharp distribution around the real value. On the other hand, a spectrum that contains pure noise will yield an even distribution of redshift values over the entire tested redshift range, and thus the average deviation from the median or mean will be quite high. In the distribution of so-called errors, which correspond to the deviation of reference redshifts across similar spectra, one would naively expect a superposition of two behaviors. The dominant component is a distribution that shows a drop toward higher deviations with a width that is comparable to the sensitivity of the method. This distribution corresponds to redshifts based on true absorption or emission lines. Underlying the first component there is a flatter distribution that represents the spectra that contain mostly noise. This is discussed further in Sect. 4.

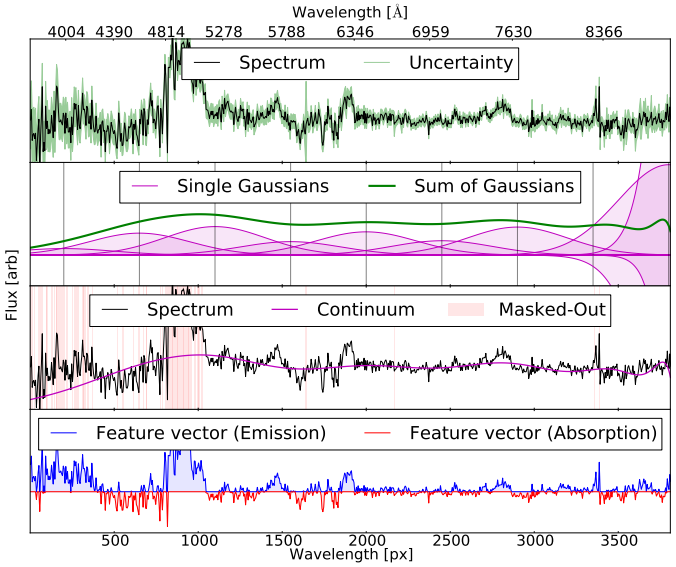


Fig. 2. *Top*: spectrum with uncertainty. *Upper center*: decomposed continuum representation by Gaussians. *Lower center*: spectrum with continuum fit and masked regions (ignored when fitting the continuum). *Bottom*: the extracted feature vectors are solely all pixel values with a value above (below) zero for emission (absorption), and all other values are set to zero.

3.3. Preprocessing

The preprocessing is needed to make the spectra comparable. Effects like apparent brightness are not important, since we are interested solely in absorption and emission features, therefore the behavior of the continuum has to be estimated and subtracted.

3.3.1. Regridding

The dispersion resolution between different fibers on a single plate and between the plates themselves differ slightly. To always be able to compare the correct wavelength bins, which do not exactly agree with redshift bins, the spectra have to be regridded. We therefore create a global grid that is defined by

$$\log(\lambda(p)) = 0.0001 \cdot p + 3.5222 \quad (4)$$

where λ is the wavelength in Å for a given pixel position p, with $0 \leq p < 5100$. The parameters of the function are chosen such that the dispersion solution corresponds to the average of our selected subsample. The regridding is performed such that the total flux is conserved.

3.3.2. Continuum estimation

The determination of the continuum is a very tricky problem that is known to cause difficulties when performing it automatically. For this reason we do not use the traditional continuum estimates (e.g., spline fitting, local weighting of polynomials[2]) and use a new hybrid method consisting of the following three approaches:

1. fit multiple Gauss model to the data,
2. weight penalty function with variance,
3. iterate three times, perform κ-clipping.

To save computation time we follow the approach by Gieseke (2011) and use a multiple Gauss decomposition via gradient based optimization. This minimizes the risk of over- or under-estimating the continuum flux, as well as over-fitting, which can be encountered when applying spline fits. To fit the continuum, a number of n normalized Gaussians with the same width w [px] are placed on the dispersion axis with the first Gaussian being placed with an offset ω [px] and all following with a spacing of d [px]. The intensity of every individual Gaussian is a free parameter to be fitted. In comparison to polynomial and spline fitting, the decomposition is less sensitive to individual spectral features and the computational effort is significantly lower. An illustration of the decomposition is shown in Fig. 2. The root-mean-square is computed based on the initial fit that is weighted with respect to the uncertainty *ivar* (see 3.3.3). Afterwards, pixel values where the difference between fit and model exceeds κ·rms are masked out for all future iterations of the continuum estimation. This is helpful for excluding large-scale deviations and accounting for detector or night-sky artifacts.

Adjacently the spectrum is now normalized with respect to the estimated continuum C by a simple min-max-normalization:

$$\mathrm{Flux}_{\mathrm{norm}} = \frac{\mathrm{Flux}_{\mathrm{raw}} - \min(C)}{\max(C) - \min(C)} \quad (5)$$

such that the continuum of the normalized flux is located between 0 and 1 and the features are normalized with respect to continuum. Since only the features are of interest for the next task the continuum is subtracted such that a flat spectrum is obtained. While we were testing different pre-processing parameters, it turned out that the quality of the overall redshift only marginally depends on the parameters used for estimating the continuum. An overview of the parameters and their impact on computation time and the quality is given in Table 1. In contrast to the literature, we treat the Ca-break also as a feature, and thus if the continuum behaves smoothly around the break, it can be seen as two close-by absorption lines afterwards.

3.3.3. Uncertainties

The SDSS spectra are affected by several uncertainties stemming from the night sky, by detector deficiency, and by read-out noise

[2] E.g., *onedspec*-package in the Image Reduction and Analysis Facility (IRAF) software package or *norm.pro* from the Interactive Data Language (IDL) software.

Table 1. Parameters used in preprocessing with tested value range and impact on the outcoming distribution as well as on the time effort for the preprocessing.

	Parameter description	Range [used value]	Impact [result / time]
n	Number of Gaussians	8–20 **[12]**	low / linear
d	Spacing between centers	300–700 **[450]**	low / none
ω	Initial center offset of first Gaussian	100–400 **[200]**	none / none
w	Gaussian width	100–1000 **[300]**	low / none
i	Number of iterations for sigma clipping	1–3 **[3]**	none / linear
κ	Noise deviation for feature refitting	0.1–3 **[0.3]**	low / none

which are quantified pixel-wise by the inverse variance *ivar*, which corresponds to the noise uncertainty σ given by

$$\sigma = \frac{1}{\sqrt{ivar}}. \tag{6}$$

After normalizing *ivar* with respect to the continuum as described above, the extracted signal-to-continuum spectra are divided by 3σ in order to normalize the noise to values between -1 and 1. Those are called normalized S/N spectra (NSN-spectra hereafter). As a consequence, the contrast between real signals and noise is increased further and artifacts stemming from a bad sky subtraction or bad pixel are heavily suppressed.

3.4. Feature extraction

To extract the feature vectors, we split the spectra into positive and negative flux components with respect to the fitted continuum (see bottom plot in Fig. 2). We thereby create two feature vectors per spectrum. This simplification allows keeping the entire redshift-dependent information while no longer being dependent on the continuum shape. The separation enables us to compute individual redshifts for absorption and emission. By extracting subregions of this feature vector, we can even obtain redshift information on single spectral regions. All values above the continuum (>0) are included in the feature vector for emission, and all values below the continuum are simply set to zero. For absorption values larger than zero are set to zero. Those extracted vectors are the input for our kNN-search described in Eq. (3).

4. Experiments

We conducted two experiments with different selections of features and different reference samples. In the following they are named Experiment 1 and 2.

4.1. Description of experiments

Both runs have been done on the full set of NSN spectra. For the first experiment, we applied the algorithm to the entire spectra and only distinguish between absorption and emission. In the second experiment we limited the dimensionality of the feature vector by just comparing specific spectral regions where features are expected for the redshift given by SDSS.

Table 2. Regions considered.

Spectral type	$\lambda_{\rm low}$ [Å]	$\lambda_{\rm high}$ [Å]	Name
Emission	2799	2799	MgII
	3346	3426	NeV
	3727	3729	[OII]
	3798	3835	H_ϵ H_ζ
	4102	4102	H_δ
	4341	4363	H_γ
	4861	5007	H_β [OIII]
	6550	6584	H_α [NII]
	6716	6731	[SII]
Absorption	3934	3969	H+K
	5173	5173	Mgb
	5890	5896	NaD

Naively, one would expect high precision in the former method because the full information content is available and thus the confusion between features of different origin (e.g., misidentifying H_β as H_α) should be fairly low. Other emission/absorption signatures are available to cross-validate the redshift and hence minimize the probability of confusion. On the other hand the obtained final regression value is only valid for the entire spectrum and thus generalizes the information content too heavily.

For this reason a second experiment was conducted with a comparison restricted to single regions where prominent emission/absorption signatures are expected. It is worth noting that this experiment is tailored to detect shifts in individual spectral lines. Additionally, the methodology can be easily extended to allow a clustering or classification of the individual lines. We assume that the redshift of SDSS is correct for the entire spectrum, but we search for redshift deviations in individual components. Since confusion will have a significant impact on the determination of the redshift, we restrict the redshift deviation of the reference sample to a spectral window W defined by

$$W = \lambda_1 - \lambda_0 = \left(1 + z_{\rm high}\right) \cdot \lambda_{\rm high} - (1 + z_{\rm low}) \cdot \lambda_{\rm low} \tag{7}$$

with

$$\begin{aligned} z_{\rm low} &= z_{\rm target} - f \cdot \left(1 + z_{\rm target}\right) \\ z_{\rm high} &= z_{\rm target} + f \cdot \left(1 + z_{\rm target}\right) \end{aligned} \tag{8}$$

where f is the allowed deviation from the SDSS redshift ($z_{\rm target}$) in units of the speed of light. A list of the spectral regions that have been considered can be found in Table 2. This list contains lines that are usually strong in star-forming and star-bursting galaxies and QSOs. The free parameter f influences the computational efforts, the chance of confusion (improving for small f) and the sensitivity to outliers where huge redshift deviations were achieved with large f, respectively. Throughout this paper, we use a value of $f = 0.05$. The big disadvantage of the second experiment is that confusion becomes a major concern. Especially for the entire data set, it might be wise not to compare spectra to spectra of any redshift since it is likely that, for example, the H_β (λ4861) line can look similar to the [OIII] (λ5007) line, see Fig. 3, which obviously would lead to an incorrect regression value. The benefit of this concept is its huge flexibility. Redshifts can now be computed for individual regions independently, so that shifts can be detected. A more detailed discussion of the trade-off of confusion and multiregion regression value determination is given in Sect. 5.

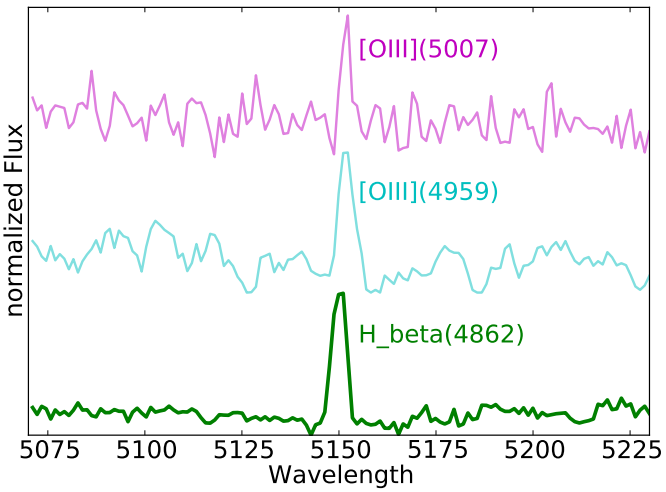


Fig. 3. Cut-out of the same spectral region for three spectra with different redshifts. If only this part is available, the H_β line is indistinguishable from any of the two [OIII] lines.

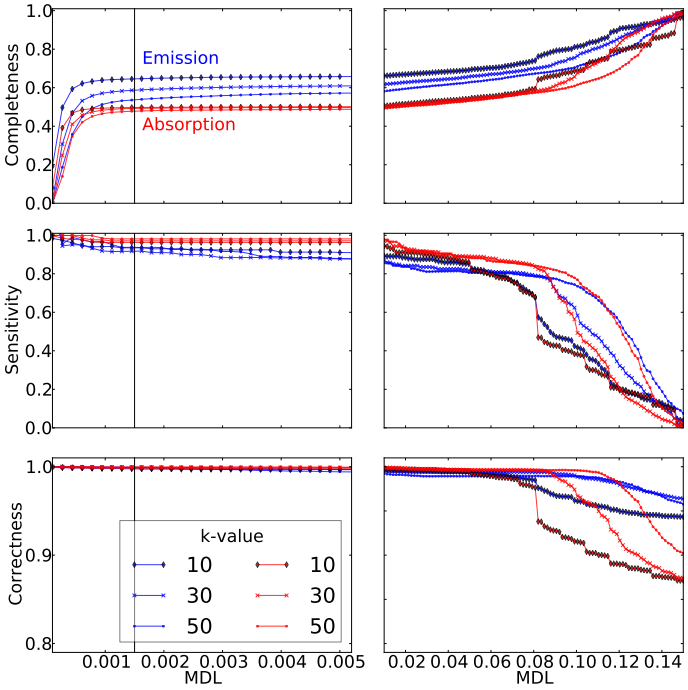


Fig. 4. Normalized completeness, sensitivity, and correctness tested against different values of k and MDL. MDL is chosen to be 0.0015 as marked.

4.2. Maximum deviation limit

One of the prerequisites in using the kNN approach was that a clear separation between noise and signal can be made. In principle there are two ways to reject spectra with no signal, the pre- and the post-selection. To preselect one assumes that a signal has a certain shape and exceeds a given S/N limit. This can be simplified further to a measure that compares the average of a spectral region with a nominal value. Because this preselection requires detailed knowledge about the shape, size and symmetry of spectral features, physical knowledge about the morphology of lines is required. To be independent of physical assumptions[3], the possibility of a post-selection is chosen. The selected concept assumes that the deviation of the redshifts of the nearest neighbors over all targets follows a smooth distribution. For this distribution an upper limit can be (freely) selected that separates redshift estimates into good and noisy ones. This maximum deviation limit will be abbreviated by MDL.

An even bigger advantage of this method is that it allows experimenting with this free parameter in the evaluation stage, such that the kNN search is not performed for every individual value of MDL.

4.3. Validation strategy

To avoid biases in the regression values and when tuning the parameters, the leave-one-out strategy is used. This means that the closest object (which is always the object itself) is not used for determining the redshift.

The fundamental assumption that most SDSS redshifts are correct has already been discussed in Sect. 3.2. Assuming now that all redshifts are correct, we can compute something like a completeness, a correctness, and a sensitivity. The completeness is a very straightforward measure. It is the fraction of objects for which a redshift could be determined within the respective acceptance limit. In contrast to that the correctness is the fraction of objects where the computed and the SDSS redshift agree within their errors. Finally, the sensitivity gives the reliability of all redshifts; i.e., it gives the typical deviation from the redshift, therefore the standard deviation of the difference between SDSS and computed redshift of all valid spectral features is computed.

4.4. Parameter tuning

Despite the parameters described in the preprocessing step only two parameters have to be fine-tuned for the regression step[4]. Those are the number of k nearest neighbors used for the comparison and the MDL that marks a spectrum as reliable. With the test strategy described in the previous section, this fine-tuning can be solved on a discrete grid (see Fig. 4). Two separate things are shown in this plot, the large scale behavior of the properties on the righthand side and on the left a zoom-in to the lowest values of the MDL.

With increasing MDL, which is equal to accepting more noisy spectral features, the properties behave just as expected; while the completeness is increasing, the sensitivity and correctness of the model are decreasing. One can further see that the completeness is a fairly flat function up to an MDL of 0.08 where it starts a more rapid, step-wise increase. The regression model breaks already down at a MDL of 0.05 where the sensitivity and correctness show a steep decrease. Since the increase of the completeness is only very tiny for high values of MDL, we now focus on the region of very tiny values of MDL.

On the small scale, the completeness strongly depends on the choice of the MDL and slight increases in the MDL yield a strong increase in completeness. Then the behavior becomes very flat, so the gain by further increasing the MDL is only marginal. When increasing the MDL 5 times, the completeness fraction increases by less than 2%. It is worth noting that the completeness depends quite heavily on k. Smaller k values result

[3] Obviously the reference values by SDSS are obtained via physical modeling.

[4] This is only partially true because different ways of computing the redshift and calculating the deviation exist. Besides the parameter tuning, one has to choose a similarity measure and to preprocess and select the features, accordingly.

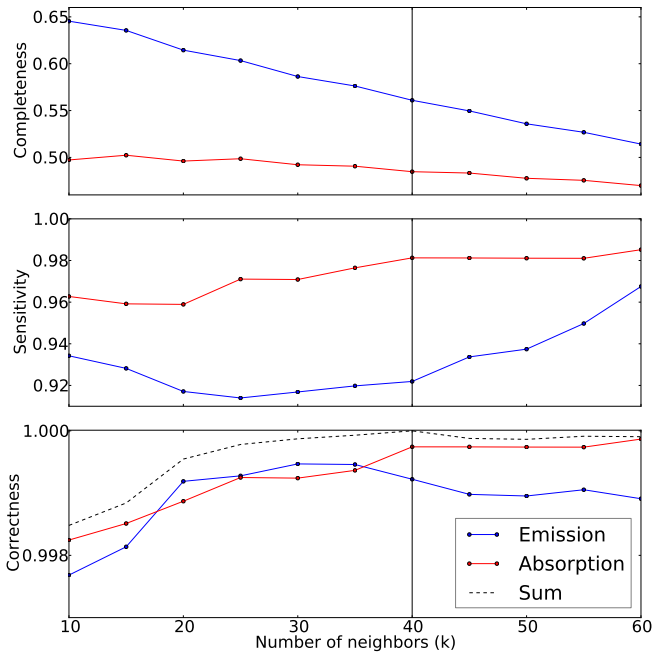


Fig. 5. Dependence of test properties on k. For this plot, MDL is fixed to 0.0015. k is chosen to be 40 as marked.

in a more complete regression model. This indicates that the number of good references is in the range of 10−20. For higher values of k more deviation is introduced in the regression model.

The sensitivity of the regression model only decreases in the beginning and then follows a fairly flat behavior with a slightly decreasing tendency. The dependence on the number of used neighbors is only marginal, though one can see that emission favors low k (little number of reference objects), while the sensitivity of the redshift in absorption is slightly better for higher values of k. In the end the fraction of outliers on the good regression side only slightly changes with increasing MDL and k. The decrease over the entire tested range is on the order of 0.5%.

The flat increase in completeness for MDL values greater than 0.001 allows us to minimize the effects on the sensitivity and correctness. We analyzed the impact of the choice of k on the different testing properties as well. The behavior of those with a fixed value of MDL of ≈0.0015 can be seen in Fig. 5. An increasing number of nearest neighbors improves the sensitivity at the cost of a lower completeness. Thus as for the MDL, the choice of k depends strongly on the desired completeness and precision.

4.5. Computational efforts

Applying the method described above to the test set is already quite time-consuming on a single machine. It is evident that the computational effort for three million spectra is many times greater than with 16 000; i.e., the time complexity of a brute force kNN search scales with $O(n^2)$, thus the calculation time would already be years on a single machine. For future surveys, this number will increase even faster such that more efficient approaches have to be found to resolve that problem. To speed up the calculation we parallelized the computation of the distances. The results presented here should only give an overview of what is possible with even the simplest methods when such a huge data amount is available.

It is worth noting that an online nearest neighbor search of incoming data (streaming) with a spectral database of the size of SDSS (≈3 000 000 spectra) is computationally feasible on a modern laptop. Assuming that a new instrument (e.g., 4MOST, de Jong et al. 2012) will obtain 2400 spectra simultaneously, the approximate comparison time is on the order of 40 h/core using a standard Python implementation. Using a machine with a simple GPU and a C-implementation will yield a speed-up of at least 100 compared to the single-CPU machine and can evaluate such a huge amount of data (<30 min) in less than the typical exposure time. Fortunately, the computation of the distances can be perfectly parallelized, so this method is well suited for larger surveys on modern computer architecture.

As already stated, the computational effort also depends strongly on the number of reference objects used for comparison, and that needs to be tuned carefully in order to minimize computing time. On the other hand, the impact of selection effects is minimized by increasing the number of reference objects that have to be chosen in the most unbiased manner.

5. Results

In the following, we use the median absolut deviation (MAD) as a deviation measure, which is defined as

$$\mathrm{MAD}_{\bar{v}} = \mathrm{median}\left(|\bar{v} - \mathrm{median}\left(\bar{v}\right)|\right)$$

for a list of values $\bar{v}$. In the following we always use the normalized difference in redshift, which is defined as

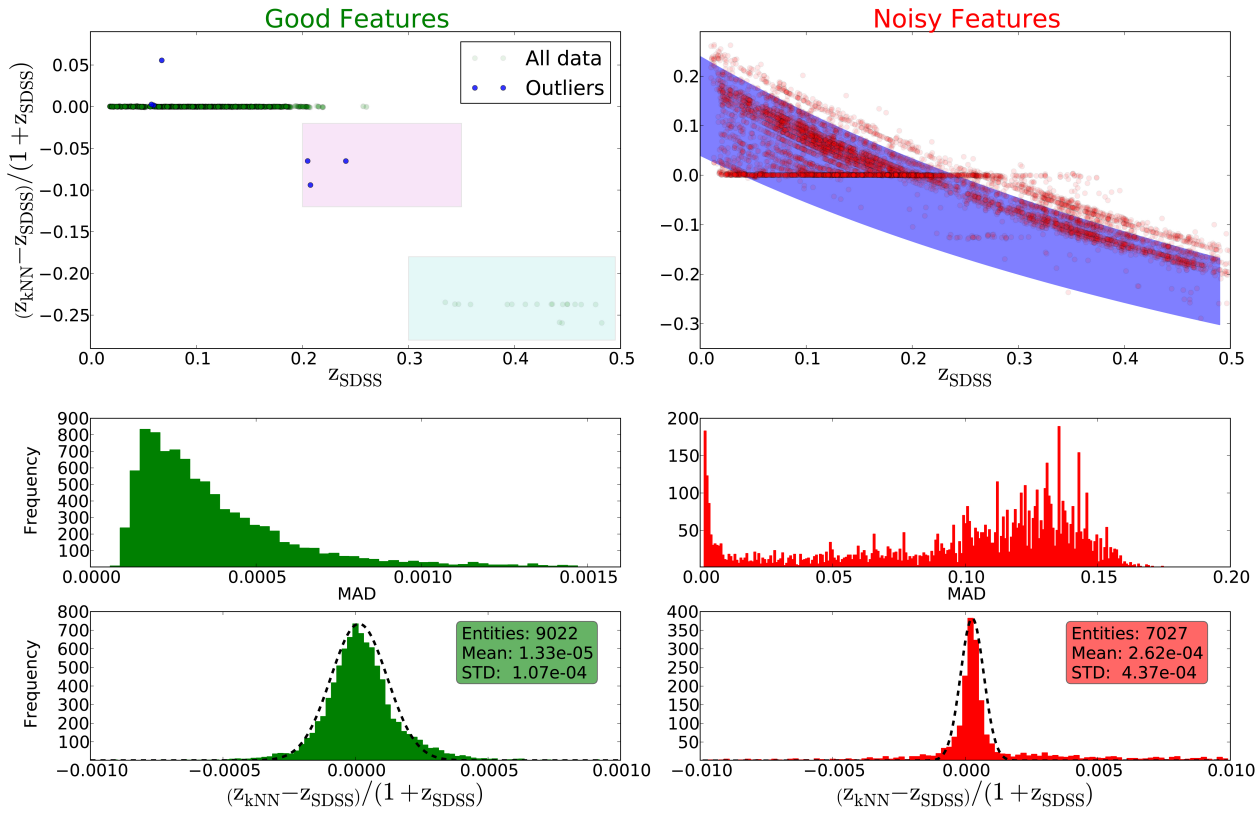
$$\Delta Z_{\mathrm{norm}} = \frac{z_{\mathrm{kNN}} - z_{\mathrm{SDSS}}}{1 + z_{\mathrm{SDSS}}}$$

which corresponds to the difference in velcocity in terms of c in the rest frame of the SDSS redshift. In Fig. 4 one can see the behavior of the completeness, sensitivity, and correctness as a function of the MDL as well as for different k. The curves follow the expected behavior; decreasing MDL will yield a low completeness but high-quality redshifts as a result. In the middle is a plateau until the MDL exceeds ≈0.10. Beyond this value the completeness starts to converge to one and the quality of the redshifts to 0, while for the value added catalog a high completeness is desirable with a moderate loss of sensitivity, so MDL = 0.07 and k = 10 are chosen. This increases the fraction of objects with a reliable redshift either in emission or in absorption up to a total of 80%. With this choice of parameters we still have better sensitivity than does SDSS with a significantly lower value in completeness (in SDSS ≈96% of the targets have NO redshift warning). As stated earlier, the choice of the reference sample, especially at high redshifts, will increase this fraction of our method significantly. For example, when we exclude spectra with $z > 0.25$ the completeness increases to ≈90%.

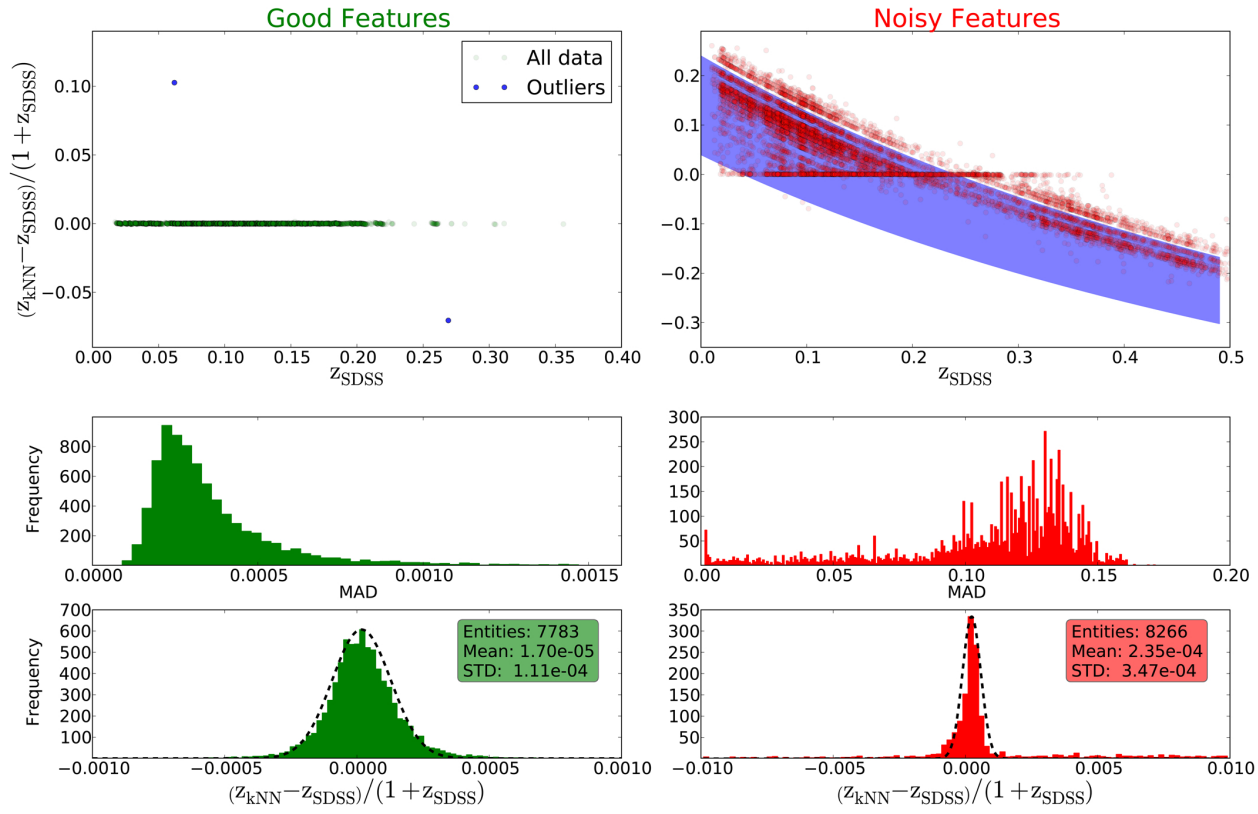
In the following we concentrate on detecting and verifying outliers using MDL = 0.015 and k = 40. With that choice we have traded high sensitivity for a lower completeness of ≥50%. This enables us to efficiently detect outliers that show wrong or multiple redshift components. In the following, the outlier detection for both experiments is discussed in detail.

5.1. Experiment 1

When using the entire spectral range for computing the redshift, we can obtain redshifts for 56% (emission) and 49% (absorption) of the spectra. Figure 6 presents the evaluation of the



(a) Emission



(b) Absorption

Fig. 6. Evaluation of the performance for emission **a)** and absorption **b)**: the relative deviation from the SDSS redshift as function of the SDSS redshift (*top*), the distribution of the MAD of the calculated redshift (*middle*) and the frequency of the relative deviation (*bottom*) are shown for the good (*left*) and the rejected noisy (*right*) spectral features, respectively. The blue background shade in the *upper right figure* reflects the objects which are entirely dominated by noise and thus their computed redshift just reflects a random draw of redshifts from the initial distribution, see Eq. (9).

achieved performance. In the second row of each figure, one can see the frequency of deviations for emission and absorption[5]. As expected, there is an exponential drop-off and an underlying uniform contribution. The top figure shows the relative deviation (in units of the speed of light) between the redshift by SDSS and the computed ones. For nearly all of the objects with prominent features, this deviation is below 0.1% c, which corresponds roughly to the SDSS resolution.

In emission one can see three groups of outliers: three points between a redshift of $0.2 \leq z \leq 0.3$ (G1, magenta background), a straight line in the lower right of the plot (G2, cyan background), and three points that deviate significantly from the expected redshift below a redshift of $z \leq 0.1$ (G3, blue dots). The cause of each of the outliers groups is different but nonetheless understood. The members of G1 are affected by the lack of reference objects in a comparable redshift range ($z > 0.2$), which agrees perfectly with the distribution shown in Fig. 1. Thus the nearest neighbors will all have a lower redshift, moving all of those points to this region in the plot. It is worth noting that one would naively expect all of those points to lie on a horizontal line, and the deviation from the reference set should be the same for all objects. In fact, it turned out that the lowest point in this group is a truly shifted object. G2 is actually a superposition of the problem just described and what was defined earlier as confusion. This confusion occurs since the relative shift in redshift of $\Delta Z_{\mathrm{norm}} \approx -0.25$ corresponds roughly to the shifts between $H_\alpha - H_\beta$ ($\Delta Z_{\mathrm{norm}} = 0.26$), H_α – [OIII] ($\Delta Z_{\mathrm{norm}} = 0.24$), [NII] – H_β ($\Delta Z_{\mathrm{norm}} = 0.26$), and [NII] – [OIII] ($\Delta Z_{\mathrm{norm}} = 0.24$). In this case the spectra usually show strong emission in either H_β or [OIII], which are then (due to missing references) misidentified as [NII] or H_α. Finally spectra with real shifts are likely to be observed close to the horizontal green line. They are discussed further in Sect. 5.3. The behavior of the noisy features can be explained by another superposition of two effects. The first group of objects is the one where the relative deviation is fairly low over the entire redshift range. Those objects are the result of the choice of the MDL – their redshift is still very accurate but they were moved to the uncertain features. A large number of spectra can be described very nicely with the applied model. This indicates that the MDL was selected quite conservatively. The rest of the data points in this plot do not show any signal of an emission feature, so they are just a random selection of redshifts from the initial distribution shown in Fig. 1. The distribution of redshifts is approximated fairly well by a Gaussian (mean = 0.14 and standard deviation = 0.10). The functional form (cf. blue background plot in upper row) is

$$((0.14 \pm 0.1) - z_{\mathrm{SDSS}}) \,/\, (1 + z_{\mathrm{SDSS}})\,. \tag{9}$$

In absorption two outliers could be detected that show some anomalies that are described well by the computed redshift. Even redshifts with high MAD are still fairly reliable, supporting the restrictive limit on the MDL. The precision in absorption is close in value to the emission, one per mil in units of the speed of light. Obviously the chance of confusion is dramatically lower than for the emission, which is the consequence of having fewer potential features. In a regular galaxy only three strong absorption features can typically be observed.

[5] Note that the bin width is changing by a factor of 25 from the left to the right side. For this reason the frequency between the two plots is not directly comparable

5.2. Experiment 2

In contrast to the first experiment the number of potential nearest neighbors of a specific spectral region now depends strongly on the choice of the redshift bin and additionally on the likelihood of the respective feature appearing in a galaxy spectra. It then becomes inevitable to discuss the chosen regions individually. To still have a good comparison of the redshifts between the different regions, the MDL is set to 0.0015. For the sake of completeness, all the figures comparing the noisy and the good features are presented in the Appendix A. Without restricting the results any further, the number of potential outliers increases drastically owing to the problem of additional confusion with different spectral features as well as to the limited number of used reference objects. Thus in order to remain clear and minimize the effect of methodological artifacts, the deviation/outlier constraint is not just tested for $k = 40$ but for a full list of nearest neighbors, namely $k = [5, 10, 20, 30]$. If the MAD violates the MDL or if the computed redshift agrees in its tolerance with the SDSS redshift for any k, the object is not marked as an outlier. Likewise, objects that have redshifts $z < 0.05$ or $z > 0.3$ are automatically excluded from the outlier detection algorithm because here the limited number of comparison objects introduces spurious redshifts. Since the different regions are biased by different effects, they are discussed in more detail.

In the following we discuss the individual spectral emission and absorption features, along with groups and individual outlying spectra. Exemplarily extensive plots for two spectral features are shown in Figs. A.1 and A.2 for H_β and NaD, respectively.

5.2.1. Emission

MgII, NeV (λ2, 799, λ3, 346-3,426). For those spectral regions a redshift of $z = 0.45/0.18$ is required to allow for a redshift determination. Since the number density of objects is fairly sparse for such high redshifts, and the NeV feature does not occur in many of those spectra, none of the redshifts can be trusted.

[OII] (λ3, 727-3, 729). This feature does not occur in all star-forming or active galaxies so that fewer than half of the redshifts could be trusted. Even in this small fraction of objects, two outliers were detected that both show an actually shifted [OII]-line that is correctly described by the value determined by us.

H_ϵ, H_ζ (λ3, 798-3, 835). One of the objects found to have a shift in the [OII]-feature could be rediscovered. Both the other additional spectral features are real.

H_δ (λ4,102). This spectral feature only appears in emission for star-forming, bursting and active galaxies. The number of reference objects exhibiting a clear sign of emission is fairly rare. In the corresponding plot one can see that two straight regions are apparent at $\Delta Z_{\mathrm{norm}} = 0.04/-0.04$, which are caused by confusion. The remaining object shows some very strong noise in the vicinity of the expected spectral feature.

H_γ,[OIII] ($\lambda 4,342 - 4,363$). The only two remaining spectra have $\Delta Z_{\mathrm{norm}} = -0.032$. When investigating the origin of this shift, it appears that the shift is dominated by noise because the number of active and starburst objects (objects that possibly emit strong Balmer lines) in the specific redshift bins is very low (<5). When selecting the redshift those few objects are therefore

strongly dominated by noise. Consequently, this feature is not very reliable as long as no representative reference sample can be selected.

H_β, [OIII] (λ4, 861-5, 007). In this spectral region the impact of confusion becomes dominant. Fourteen objects show a reasonably low deviation to be marked as good estimates. The horizontal line at $\Delta Z_{\rm norm} = 0.03$ is caused by a misidentification of the red [OIII] line with the H_β-feature. The line at $\Delta Z_{\rm norm} = 0.01$ is due to the confusion between the red and the blue [OIII] lines. The negative confusion at $\Delta Z_{\rm norm} = -0.03$ is the reverse effect of the first one. Another horizontal component at $\Delta Z_{\rm norm} = -0.055$ is caused by a misidentification between the blue [OIII] line and HeII emission at λ4,685.

Apart from all this confusion there is one regular shift that cannot be confirmed owing to the lack of other emission features. The MAD for this object (0.0014) is close to the MDL so a lower choice of the MDL would tag this object as unreliable.

H_α, [NII] (λ6, 550-6, 584). The outlier on the very top of the plot was already marked by the first run and is a truly shifted spectral feature. One of the shifts of the remaining two outliers is the result of an H_α-line in absorption and emission such that the red [NII]-line was mistaken for it. In the other a very weak [NII] emission line led to confusion with the H_α line.

[SII] (λ6, 716-6, 731). The only object marked in the plot was also detected in the H_α-line as an outlier. It was already marked as an outlier in Experiment 1.

5.2.2. Absorption

CaII (HK) (λ3, 934-3, 969). All three targets highlighted as outlier are truly shifted spectral features, where one of them is the object already detected in emission (cf. Experiment 1).

Mgb (λ5, 173). Six of the objects are located on a horizontal line around $\Delta Z_{\rm norm} = -0.06$. This corresponds to a misidentification of the Mgb absorption with the H_β in absorption. Indeed, all highlighted objects show a very prominent H_β feature in absorption. Two of the remaining objects have a very strong absorption feature stemming from deficient nightsky subtraction that was not properly described by *ivar*. Three objects are active galaxies, and they show extremely strong emission features in this region. The number of active galaxies in the reference sample is not sufficient to reproduce this behavior. The remaining object shows a true shift in the Mgb line.

NaD (λ5, 890-5, 896). For seven spectra a shift of the NaD could be confirmed by a manual inspection. For all the others a badly subtracted sky at around λ7200 was not described correctly by *ivar*, leading to a very prominent absorption feature that was mistaken for NaD.

5.3. Manually investigated objects

To validate the method, a manual inspection of the outliers is mandatory. A spectrum was investigated if it was selected as an outlier in any of the spectral regions (from Experiments 1 and 2) and if it was not part of one of the horizontal lines introduced by confusion. The outliers have different origins that can be roughly classified into three groups: objects with real multiple redshift components (true), objects with detectoror nightsky artifacts that were not properly described by *ivar* (fake), and objects where the redshift computation simply failed (wrong).

Thirty-seven objects were eventually investigated manually, three of which have been marked by several features as outliers. Fourteen of the outliers (38%) are spectra with truly shifted redshift components. In 11 of those, the shift between the redshift components is lower than 10 000 km s^{-1}, so those components certainly do have a physical origin. The three remaining spectra of the true class are likely to be superpositions and/or lensed objects. The fake category contains ten objects where a badly described detector/nightsky artifact was confused with NaD or Mgb absorption. It is impossible to exclude those objects previously because there is no unique position or indication of the existence of such a feature. The thirteen spectra in the wrong class are mainly the result of a biased reference sample that also contains a low number of active and star-bursting galaxies. There is a good chance that the fraction of those objects can be significantly decreased if a more representative reference sample is used for the comparison.

A short summary of all manually investigated objects with identifier, SDSS, and computed redshift can be found in Table A.1.

5.4. Most prominent outliers

The most prominent outliers are briefly described here to emphasize the power of this outlier detection scheme. In Fig. 7 one can see the three truly shifted objects with the highest velocity offset. While the first two (J094419.05-004051.44, J120419.07-001855.93) were even tagged independently by the separate runs, the last one (J113154.29+001719.02) did not show up in the second experiment because the relative shift between our computed and the SDSS redshift (0.077 c) exceeds the allowed range of the shift (0.060 c). In the first and last objects the model applied by SDSS describes the absorption behavior quite well, but the emission features are not described at all such that a second component with a strongly shifted redshift is needed to describe them. While they demonstrate the power of the method quite nicely these objects are astronomically less interesting. Owing to missing signs of interaction, it is likely that they are just simple superpositions of objects. In the *i*-band of the first object, a tiny and asymmetric arc (cf. Fig. 8) can be seen that could indicate a lensed object. The redshift of the second object was estimated incorrectly by SDSS because apparently none of the template models was able to describe both the continuum and the line behavior at the same time. The newly estimated redshift, on the other hand, describes the spectrum quite well. While the new fit does not support the existence of another component, it is worth noting that on the SDSS image a clear symmetric arc can be seen at a distance of a few arcseconds.

5.5. Summary of outliers

In Table A.1 all outliers found are summarized. Targets where a true feature exists are marked. The possible origins for the existence of multiple redshift components are miscellaneous. For a very high shift between the redshifts, the most likely explanation is a chance superposition of two objects. In this case an

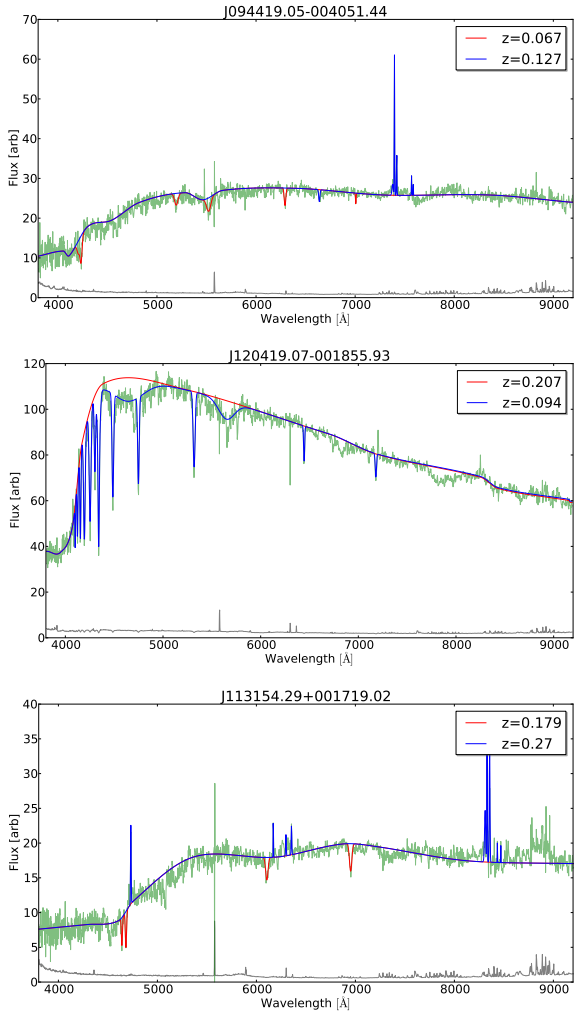


Fig. 7. Three most extreme outliers obtained from our regression model. The green line in the background is the SDSS spectrum with the gray spectrum at the *bottom* being the typical noise deviation. The red curve shows the fitted spectrum with a redshift as obtained by SDSS, the blue curve is the overplotted spectrum with the redshift as obtained by our regression model.

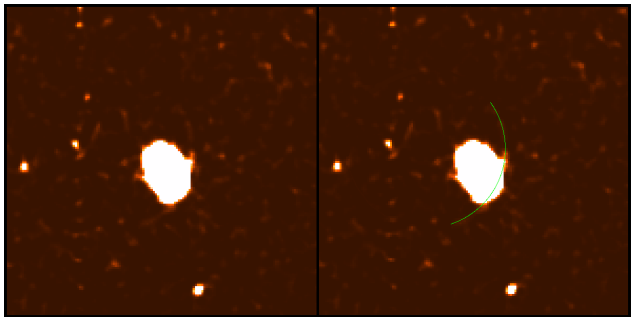


Fig. 8. SDSS *i*-band image of J094419.05-004051.44 (smoothed with Gaussian blur of 3-pixel width). In the *right image* the asymmetric arc has been overplotted by a green circle.

arc due to gravitational lensing might be observed. The number of gravitational lenses in the near universe is very limited so far (Muñoz et al. 1998). Spectra with the velocity shifts between the lines lower than $<10\,000\,\mathrm{km\,s^{-1}}$ might be good candidates for being super-massive black hole binaries (SMBHB, Tsalmantza et al. 2011; Fu et al. 2012; Popović 2012). The kinematics of the broad line region are a very common cause of such observed line shifts as well (Shen et al. 2011).

Eventually one could only distinguish between the different origins by either deep-imaging (lenses) or follow-up spectroscopy (SMBHB, Liu et al. 2014). High-resolution imaging in the multiple wavelengths could also distinguish single from multiple sources (e.g., Rodriguez et al. 2009).

6. Summary

This paper presents a new methodology that performs a redshift computation based on pre-exisiting SDSS redshifts. The aim is to obtain improved redshifts for emission and absorption, as well as for individual spectral features. This enables astronomers to detect spectra with multiple redshift components. The basic principle of the presented method is to perform a self-consistency check such that objects that look similar should have a comparable redshift.

First of all, it is worth noting that this method performs quite well in calculating the redshift for very different kinds of spectra. The only requirement is that the density of reference objects is reasonably high in the d-dimensional Euclidean space populated by the spectra. It could be shown that with its current set of reference spectra (which is limited to redshifts $z \leq 0.5$, but the reference sample is just densely populated until $z \approx 0.2$) this method can reach higher sensitivity than the SDSS pipeline for individual spectra. So far, only the completeness is considerably lower than in the SDSS pipeline, but this will be improved using a larger and more representative reference sample that covers all redshifts.

To show the power of this new tool, we presented outliers found in the data set. For this a more conservative (more sensitive, but less complete) parameter set was chosen. We were able to detect outliers by two different statistical redshifts. The first approach focuses on the overall behavior of the spectra, so is less affected by confusion but is less informative. The second approach focuses on the behavior of predefined regions. Its completeness rate is higher; i.e., more objects with exotic behavior have been found. On the other hand, the number of highlighted objects that appear due to methodological artifacts is also increased. In summary both methods yield very interesting objects where the SDSS redshift was incorrect.

Even though these methods work quite well, plenty of parameters exist that are tunable and that have an impact on the final result. In the data preprocessing several models describing the continuum behavior were investigated. The normalization of the spectra with respect to this continuum and their noise might have an effect on the number of true outliers, too. In addition, the feature extraction has a strong impact on the final results and might be tailored to certain scientific needs.

In a next step, we will investigate the impact of the choice of the reference sample. Each redshift bin should contain enough reference objects to minimize systematic effects of to the sample bias. This discussion is part of a forthcoming paper, where the methodology is applied to the full SDSS spectroscopic database.

In a final step, the impact of the mathematical composition of the regression values used in Eq. (3) could be investigated. It would also be interesting to study the behavior of different selection measures, such that a clearer distinction between noisy and good features can be made. Additionally, one could apply a pre- instead of a post-selection to distinguish between signals and noise on the data level. This would make the reduction of the

reference sample in the computational step easier, as only reference objects with an existing signal would be used for comparison. On the other hand, it would introduce more biases that have to be tuned by the increased number of parameters. Some physical knowledge about the type of signal that is expected would be required.

Finally, the outlier detection could be modified. Depending on the scientific use case, the trade-off between completeness and sensitivity can be adjusted by using different detection criteria. Those detected outliers can be related to future outlier catalogs. Since we are currently only investigating a small fraction of the database (<1%), a huge number of objects are expected to be marked as outliers for the entire dataset; i.e., that the number of objects to be investigated will be so large (≈5000) that a manual inspection will be extremely time-consuming. At any rate, the discovery potential of this straightforward redshift determination approach is huge. The applicability on new incoming data was already shown on this simplified and just partially representative sub-sample.

Acknowledgements. Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the US Department of Energy Office of Science. The SDSS-III web site is http://www.sdss3.org/. SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofisica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University. The authors thank Michael Schick for very fruitful discussions on the topic of uncertainty quantification. S.D.K. would like to thank the Klaus Tschira Foundation for their financial support.

Appendix A

Table A.1. Summary of all manually investigated spectra.

Identifier [plate-MJD-fiber]	z_{SDSS}	z_{EST}	Δz [km s^{-1}]	Spectral feature	Class	Remarks
0266-51 602-0095	0.0673	0.1266 ± 0.0005	16 600	Em,H$_\alpha$,[NII],[SII]	true	Fig. 7 top
0286-51 999-0236	0.2075	0.0940 ± 0.0007	–28 200	Em,CaII,NaD	true	possible lense, Fig. 7 center
0268-51 633-0423	0.1699	0.1689 ± 0.0005	–300	OII,H$_\epsilon$,H$_\zeta$	true	shifted BL/NL
0275-51 910-0265	0.0596	0.0614 ± 0.0013	500	Em	wrong	low number density of reference objects (see Fig. 1)
0279-51 984-0449	0.2049	0.1264 ± 0.0014	–19 600	Em	wrong	member of G1
0280-51 612-0323	0.0576	0.0605 ± 0.0007	800	Em	wrong	low number density of reference objects (see Fig. 1)
0282-51 658-0493	0.2409	0.1600 ± 0.0009	–19 600	Em	wrong	member of G1
0267-51 608-0601	0.0620	0.1709 ± 0.0006	30 700	Abs	fake	fake feature at λ6, 901
0282-51 630-0400	0.2690	0.1796 ± 0.0008	–21 200	Abs	true	Fig. 7 bottom
0274-51 913-0617	0.0966	0.1159 ± 0.0004	5200	OII	true	dual core in image
0272-51 941-0332	0.2201	0.2184 ± 0.0010	–500	H$_\epsilon$,H$_\zeta$	true	NL shifted vs. absorption
0288-52 000-0215	0.1531	0.1543 ± 0.0007	200	H$_\epsilon$,H$_\zeta$	true	shifted BL/NL
0268-51 633-0354	0.0918	0.0928 ± 0.0006	200	H$_\delta$	wrong	strong noise in spectral region
0271-51 883-0371	0.1202	0.0851 ± 0.0012	–9500	H$_\gamma$	wrong	litte active & starburst spectra
0273-51 957-0579	0.1312	0.0946 ± 0.0007	–9800	H$_\gamma$	wrong	litte active & starburst spectra
0279-51 608-0034	0.1010	0.0975 ± 0.0014	–1000	H$_\beta$,[OIII]	wrong	high MAD
0271-51 883-0570	0.0531	0.0638 ± 0.0013	3000	H$_\alpha$,[NII]	wrong	low number density of reference objects (see Fig. 1)
0286-51 999-0089	0.1296	0.1263 ± 0.0004	–900	H$_\alpha$,[NII]	wrong	very weak features only
0267-51 608-0593	0.1631	0.1642 ± 0.0005	200	CaII	true	
0275-51 910-0142	0.1516	0.1526 ± 0.0005	200	CaII	true	
0266-51 602-0604	0.2995	0.3353 ± 0.0012	8200	Mgb	fake	λ6, 913
0266-51 630-0374	0.1661	0.1449 ± 0.0007	–5500	Mgb	fake	λ5, 892
0270-51 909-0537	0.1774	0.1589 ± 0.0010	–4800	Mgb	wrong	QSO, sparse in reference
0277-51 908-0277	0.2822	0.2593 ± 0.0013	–5400	Mgb	wrong	QSO, sparse in reference
0285-51 930-0170	0.1794	0.1685 ± 0.0009	–2800	Mgb	wrong	QSO, sparse in reference
0288-52 000-0215	0.1531	0.1544 ± 0.0006	300	Mgb	true	shifted BL/NL
0266-51 630-0318	0.1483	0.1706 ± 0.0008	5800	NaD	fake	λ6, 901
0267-51 608-0092	0.1467	0.1456 ± 0.0005	–300	NaD	true	only NaD shifted
0267-51 608-0320	0.1164	0.1836 ± 0.0010	18 000	NaD	fake	λ6, 976
0269-51 910-0531	0.1278	0.1958 ± 0.0011	18 000	NaD	fake	λ7, 045
0270-51 909-0114	0.1970	0.1960 ± 0.0007	–300	NaD	true	only NaD shifted
0274-51 913-0548	0.1000	0.0979 ± 0.0002	–600	NaD	true	only NaD shifted
0283-51 660-0602	0.1281	0.1968 ± 0.0007	18 200	NaD	fake	λ7, 053
0284-51 943-0531	0.1578	0.1958 ± 0.0005	9800	NaD	fake	λ7, 048
0284-51 943-0603	0.2013	0.1985 ± 0.0006	–700	NaD	fake	λ7, 062
0285-51 663-0602	0.2426	0.1979 ± 0.0007	–10 800	NaD	fake	λ7, 058
0285-51 930-0035	0.1222	0.1212 ± 0.0002	–300	NaD	true	only NaD shifted

Notes. The spectral features Em, Abs are from Experiment 1, all others mark the respective regions where the feature was detected as an outlier. The sorting is by number of spectral features, spectral feature, and finally identifier.

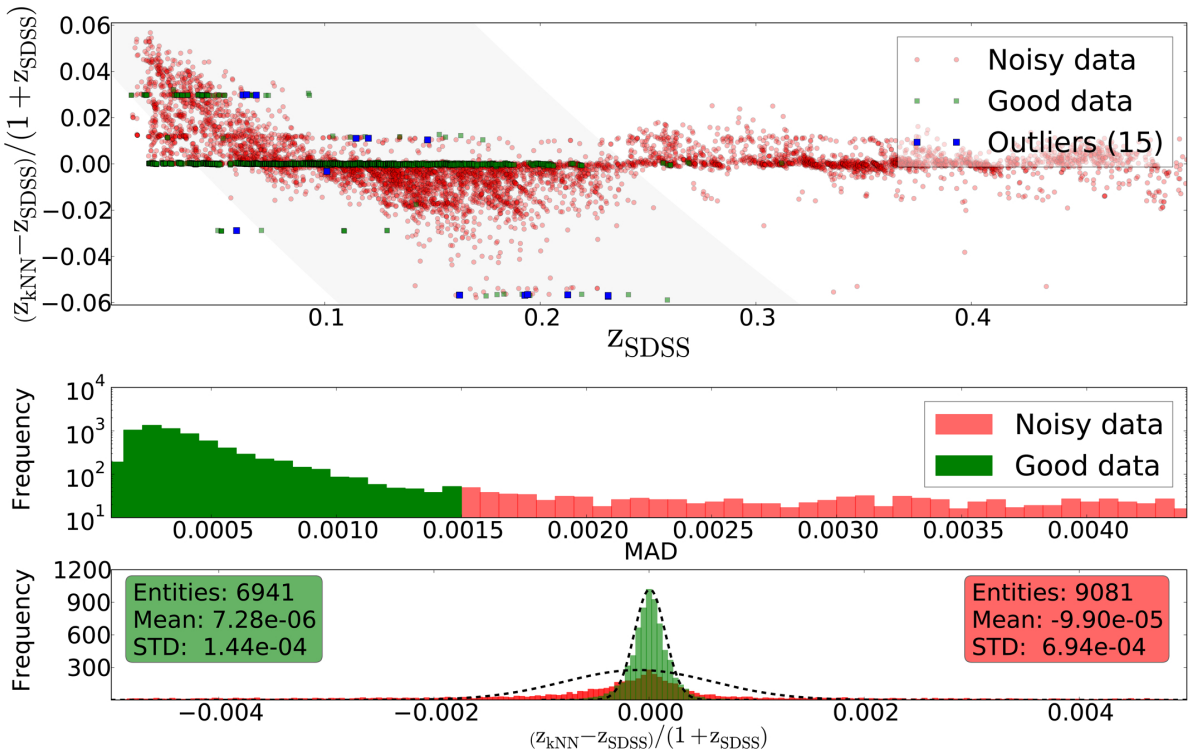


Fig. A.1. Analysis of the H_β,[OIII] region. One can see the relative difference in redshift against the SDSS redshift (*top*), the distribution of the deviations (*middle*), and a histogram of the relative difference in redshift (*bottom*). Noisy spectral features are marked in red, good features in green.

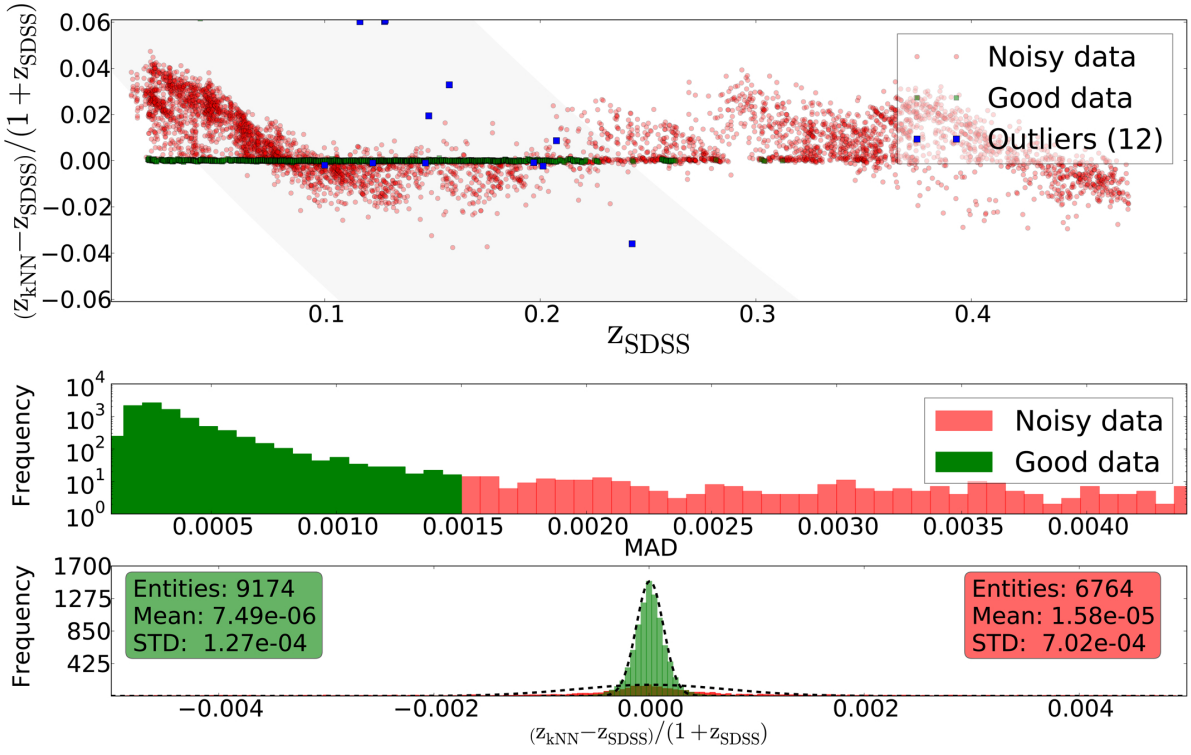


Fig. A.2. Analysis of the NaD region. One can see the relative difference in redshift against the SDSS redshift (*top*), the distribution of the deviations (*middle*), and a histogram of the relative difference in redshift (*bottom*). Noisy spectral features are marked in red, good features in green.

References

Ahn, C. P., Alexandroff, R., Allende Prieto, C., et al. 2014, ApJS, 211, 17
Ball, N. M., & Brunner, R. J. 2010, Int. J. Mod. Phys. D, 19, 1049
Bellman, R., & Bellman, R. E. 1961, Adaptive Control Processes: A Guided Tour (Princeton University Press)
Bolton, A. S., Burles, S., Schlegel, D. J., Eisenstein, D. J., & Brinkmann, J. 2004, AJ, 127, 1860
Bolton, A. S., Schlegel, D. J., Aubourg, É., et al. 2012, AJ, 144, 144
Borne, K. 2009, ArXiv e-prints [arXiv:0911.0505]
Cui, X.-Q., Zhao, Y.-H., Chu, Y.-Q., et al. 2012, RA&A, 12, 1197
de Jong, R. S., Bellido-Tirado, O., Chiappini, C., et al. 2012, in SPIE Conf. Ser., 8446
Eisenstein, D. J., Annis, J., Gunn, J. E., et al. 2001, AJ, 122, 2267
Fu, H., Yan, L., Myers, A. D., et al. 2012, ApJ, 745, 67
Gieseke, F. 2011, dissertation, Universität Oldenburg
Gieseke, F., Polsterer, K. L., Thom, A., et al. 2011, ArXiv e-prints [arXiv:1108.4696]
Hastie, T., Tibshirani, R., & Friedman, J. 2009, The Elements of Statistical Learning: Data Mining, Inference, and Prediction., 2nd edn. (Springer)
Laurino, O., D'Abrusco, R., Longo, G., & Riccio, G. 2011, MNRAS, 418, 2165
Liu, X., Shen, Y., Bian, F., Loeb, A., & Tremaine, S. 2014, ApJ, 789, 140
Meusinger, H., Schalldach, P., Scholz, R.-D., et al. 2012, A&A, 541, A77
Muñoz, J. A., Falco, E. E., Kochanek, C. S., et al. 1998, Ap&SS, 263, 51
Polsterer, K. L., Zinn, P.-C., & Gieseke, F. 2013, MNRAS, 428, 226
Popović, L. Č. 2012, New Astron. Rev., 56, 74
Richards, G. T., Fan, X., Newberg, H. J., et al. 2002, AJ, 123, 2945
Richards, J. W., Freeman, P. E., Lee, A. B., & Schafer, C. M. 2009, ApJ, 691, 32
Rodriguez, C., Taylor, G. B., Zavala, R. T., Pihlström, Y. M., & Peck, A. B. 2009, ApJ, 697, 37
Shen, Y., Richards, G. T., Strauss, M. A., et al. 2011, ApJS, 194, 45
Strauss, M. A., Weinberg, D. H., Lupton, R. H., et al. 2002, AJ, 124, 1810
Tsalmantza, P., Decarli, R., Dotti, M., & Hogg, D. W. 2011, ApJ, 738, 20
York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, AJ, 120, 1579

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

Autoencoding Time Series for Visualisation

Nikolaos Gianniotis[1], Dennis Kügler[1], Peter Tiňo[2], Kai Polsterer[1] and Ranjeev Misra[3]

1- Astroinformatics - Heidelberg Institute of Theoretical Studies
Schloss-Wolfsbrunnenweg 35 D-69118 Heidelberg - Germany

2 - School of Computer Science - The University of Birmingham
Birmingham B15 2TT - UK

3 - Inter-University Center for Astronomy and Astrophysics
Post Bag 4, Ganeshkhind, Pune-411007 - India

Abstract. We present an algorithm for the visualisation of time series. To that end we employ echo state networks to convert time series into a suitable vector representation which is capable of capturing the latent dynamics of the time series. Subsequently, the obtained vector representations are put through an autoencoder and the visualisation is constructed using the activations of the "bottleneck". The crux of the work lies with defining an objective function that quantifies the reconstruction error of these representations in a principled manner. We demonstrate the method on synthetic and real data.

1 Introduction

Time series are often considered a challenging data type to handle in machine learning tasks. Their variable-length nature has forced the derivation of feature vectors that capture various characteristics, e.g. [1]. However, it is unclear how well such (often handcrafted) features express the potentially complex latent dynamics of time series. Time series exhibit long-term dependencies which must be taken into account when comparing two time series for similarity. This temporal nature makes the use of common designs, e.g. RBF kernels, problematic. Hence, more attentive algorithmic designs are needed and indeed in classification scenarios there have been works [2, 3, 4] that successfully account for the particular nature of time series.

In this work we are interested in visualising time series. We propose a fixed-length vector representation for representing sequences that is based on the Echo State Network (ESN) [5] architecture. The great advantage of ESNs is the fact that the hidden part, the reservoir of nodes, is fixed and only the readout weights need to be trained. In this work, we take the view that the readout weight vector is a good and comprehensive representation for a time series.

In a second stage, we employ an autoencoder [6] that reduces the dimensionality of the readout weight vectors. However, employing the usual L_2 objective function for measuring reconstruction is inappropriate. What we are really interested in is not how well the readout weight vectors are reconstructed in the L_2 sense, but how well each reconstructed readout weight vector can still reproduce its respective time series when plugged back to the *same, fixed* ESN reservoir. To that end, we introduce a more suitable objective function for measuring the reconstruction quality of the autoencoder.

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

2 Echo state network cost function

An ESN is a recurrent discrete-time neural network. It processes time series composed by a sequence of observations which we denote by[1] $\boldsymbol{y} = (y(1), y(2), \ldots, y(T))$. The task of the ESN is given $y(t)$ as an input to predict $y(t+1)$. An ESN is typically formulated using the following two equations:

$$\boldsymbol{x}(t+1) = h(\boldsymbol{u}\boldsymbol{x}(t) + \boldsymbol{v}y(t)) \ , \quad (1)$$

$$y(t+1) = \boldsymbol{w}\boldsymbol{x}(t+1) \ , \quad (2)$$

where $\boldsymbol{v} \in R^{N\times 1}$ is the input weight, $\boldsymbol{x}(t) = [x_1, \ldots, x_N] \in R^{N\times 1}$ are the latent state activations of the reservoir, $\boldsymbol{u} \in R^{N\times N}$ the weights of the reservoir units, $\boldsymbol{w} \in R^{N\times 1}$ the readout weights[2]. N is the number of hidden reservoir units. Function $h(\cdot)$ is a nonlinear function, e.g. tanh, applied element-wise. According to ESN methodology [5] parameters $\boldsymbol{v}$ and $\boldsymbol{u}$ in Eq. (1) are randomly generated and fixed. The only trainable parameters are the readout weights $\boldsymbol{w}$ in Eq. (2).

Training involves feeding at each time step t an observation $y(t)$ and recording the resulting activations $\boldsymbol{x}(t)$ row-wise into a matrix $\boldsymbol{X}$. Typically, some initial observations are dismissed in order to "washout" [5] the initial arbitrary reservoir state. The following objective function ℓ is minimised:

$$\ell(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|^2 + \frac{1}{2}\lambda^2\|\boldsymbol{w}\|^2 \ , \quad (3)$$

where λ is a regularisation term. How well vector $\boldsymbol{w}$ models $\boldsymbol{y}$ with respect to the fixed reservoir is measured by objective ℓ. The optimal solution is $\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda^2\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ where $\boldsymbol{I}$ is the identity matrix. We express this as a function g that maps time series to readout weights:

$$g(\boldsymbol{y}) = (\boldsymbol{X}^T\boldsymbol{X} + \lambda^2\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{w} \ . \quad (4)$$

3 Vector representation for time series

Given a fixed ESN reservoir, for each time series in the dataset we determine its best readout weight vector and take it to be its new representation *with respect to this reservoir*.

3.1 ESN reservoir construction

Typically, parameters $\boldsymbol{v}$ and $\boldsymbol{u}$ in Eq. (1) are set stochastically [5]. To eliminate dependence on random initial conditions when constructing the ESN reservoir, we strictly follow the deterministic scheme[3] in [7]. Accordingly, we fix the topology of the reservoir by organising the reservoir units in a cycle using the *same*

[1] For brevity we assume univariate time series, i.e. $y(t) \in R$.

[2] Bias terms can be subsumed into weight vectors $\boldsymbol{v}$ and $\boldsymbol{u}$ but are ignored here for brevity.

[3] We stress that our algorithm is not dependent on this deterministic scheme for constructing ESNs; in fact it also works with the "standard" stochastically constructed ESN type as in [5].

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

coupling weight a. Similarly, all elements in $\boldsymbol{v}$ are assigned the same absolute value $b > 0$ with signs determined by an aperiodic sequence as specified in [7]. Further, following this methodology we determine values for a and b by cross-validation. The combination a, b with the lowest test error is used to instantiate the ESN reservoir that subsequently encodes the time series as readout weights.

3.2 Encoding time series as readout weights

Given the fixed reservoir, specified by a and b, we encode each time series $\boldsymbol{y}_j$ in the dataset by the readout weights $\boldsymbol{w}_j$ using function $g(\boldsymbol{y}_j) = \boldsymbol{w}_j$ (see Eq. (4)). We emphasise that *all* time series $\boldsymbol{y}_j$ are encoded with respect to the *same* fixed reservoir. Hence dataset $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_J\}$ is now replaced by $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_J\}$.

4 Autoencoding with respect to the fixed reservoir

The autoencoder [6] learns an identity mapping by training on targets identical to the inputs. Learning is restricted by the bottleneck that forces the autoencoder to reduce the dimensionality of the inputs, and hence the output is only an approximate reconstruction of the input. By setting the number of neurons in the bottleneck to two, the bottleneck activations can be interpreted as two-dimensional projection coordinates $\boldsymbol{z} \in R^2$ and used for visualisation.

The autoencoder is the composition of an encoding f_{enc} and a decoding f_{dec} function. Encoding maps inputs to coordinates, $f_{enc}(\boldsymbol{w}) = \boldsymbol{z}$, while decoding approximately maps coordinates back to inputs, $f_{dec}(\boldsymbol{z}) = \tilde{\boldsymbol{w}}$. The complete autoencoder is a function $f(\boldsymbol{w}; \boldsymbol{\theta}) = f_{dec}(f_{enc}(\boldsymbol{w})) = \tilde{\boldsymbol{w}}$, where $\boldsymbol{\theta}$ are the weights of the autoencoder trained by backpropagation.

4.1 Training mode

Typically, training the autoencoder involves minimising the L^2 norm between inputs and reconstructions over the weights $\boldsymbol{\theta}$:

$$\frac{1}{2}\sum_{j=1}^{J} \|f(\boldsymbol{w}_j; \boldsymbol{\theta}) - \boldsymbol{w}_j\|^2. \tag{5}$$

However, this objective measures only how good the reconstructions $\tilde{\boldsymbol{w}}_j$ are in the L_2 sense. What we are really interested in is how well the reconstructed weights $\tilde{\boldsymbol{w}}_j$ are still a good readout weight vector when plugged back to the fixed reservoir. This is actually what the objective function ℓ in Eq. (3) measures. This calls for a modification in the objective function Eq. (5) of the autoencoder:

$$\frac{1}{2}\sum_{j=1}^{J} \ell_j(f(\boldsymbol{w}_j; \boldsymbol{\theta})) = \frac{1}{2}\sum_{j=1}^{J} \|\boldsymbol{X}_j f(\boldsymbol{w}_j; \boldsymbol{\theta}) - \boldsymbol{y}_j\|^2 + \frac{1}{2}\lambda^2 \|f(\boldsymbol{w}_j; \boldsymbol{\theta})\|^2 \ , \tag{6}$$

where ℓ_j and $\boldsymbol{X}_j$ are the objective function and hidden state activations associated with time series $\boldsymbol{y}_j$ (see Eq. (3)). The weights $\boldsymbol{\theta}$ of the autoencoder can now be trained via backpropagation using the modified objective in Eq. (6).

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

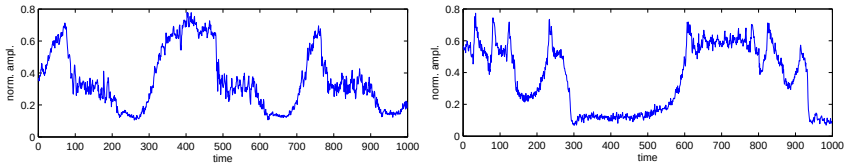


Fig. 1: Example X-ray radiation regimes β (left) and κ (right).

4.2 Testing mode

Having trained the autoencoder f via backpropagation, we would like to project new incoming time series $\boldsymbol{y}^*$ to coordinates $\boldsymbol{z}^*$. To that end we first use the fixed ESN reservoir to encode the time series as a readout weight vector $g(\boldsymbol{y}^*) = \boldsymbol{w}^*$ (see Eq. (4)). The readout weight vector $\boldsymbol{w}^*$ can then be projected using the encoding part of the autoencoder to obtain the projection $f_{enc}(\boldsymbol{w}^*) = \boldsymbol{z}^*$.

5 Experiments and Results

We present results on two synthetic datasets and on a real astronomical dataset. In all experiments we constructed the ESN reservoir deterministically according to [7] and fixed the size of the reservoir to $N = 200$. We used a washout period of 200 observations. Regularisation parameter λ for the ESNs was fixed to 10^{-4}. The number of neurons in the hidden layers of the autoencoder was set to 10. The proposed algorithm can handle out-of-sample data and hence apart from projecting training data only, we also project unseen test data. We apply no normalisation to the datasets. Moreover, we also constructed visualisations using the popular t-SNE algorithm [8] on the raw signals. We found the visualisation produced by t-SNE did not differ greatly over a range of perplexities $\{5, 10, \ldots, 50\}$.

NARMA: We generated sequences from the following NARMA classes [7] of order $10, 20, 30$, of length 800, using the following equations respectively:

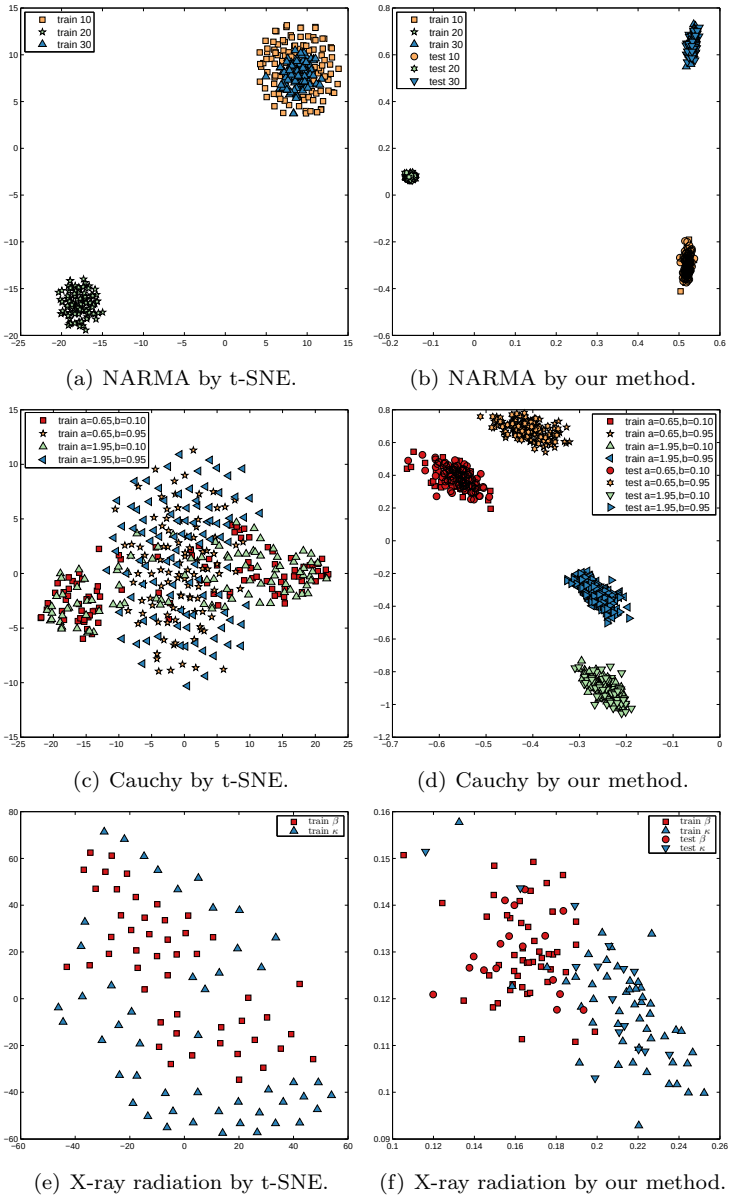
$$y(t+1) = 0.3y(t) + 0.05y(t)\sum_{i=0}^{9} y(t-i) + 1.5s(t-9)s(t) + 0.1,$$

$$y(t+1) = \tanh(0.3y(t) + 0.05y(t)\sum_{i=0}^{19} y(t-i) + 1.5s(t-19)s(t) + 0.01) + 0.2,$$

$$y(t+1) = 0.2y(t) + 0.004s(t)\sum_{i=0}^{29} y(t-i) + 1.5s(t-29)s(t) + 0.201,$$

where $s(t)$ are exogenous inputs generated independently and uniformly in the interval $[0, 0.5)$. These time series constitute an interesting synthetic example due to the long-term dependencies they exhibit.

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.



(a) NARMA by t-SNE. (b) NARMA by our method.

(c) Cauchy by t-SNE. (d) Cauchy by our method.

(e) X-ray radiation by t-SNE. (f) X-ray radiation by our method.

Fig. 2: Colours represent classes. The proposed algorithm supports out-of-sample visualisation, hence markers • and ▲ are the projections of the training and testing data respectively. Note that in the NARMA and Cauchy plots • and ▲ heavily overlap.

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

Cauchy class: We sampled sequences from a stationary Gaussian process with correlation function given by $c(x_t, x_{t+h}) = (1 + |h|^a)^{-\frac{a}{b}}$ [9]. We generated 4 classes of such time series by the permutation of parameters $a \in \{0.65, 1.95\}$ and $b \in \{0.1, 0.95\}$. We generated from each class 100 time series of length 2000.
X-ray radiation from black hole binary: We used data from [10] concerning a black hole binary system that expresses various types of temporal regimes which vary over a wide range of time scales. We extracted subsequences of length 1000 from regimes β and κ that were chosen on account of their similarity (see Fig. 1).

6 Discussion and Conclusion

We show the visualisations in Fig. 2. Unlike t-SNE which operates directly on the raw data, the proposed algorithm can capture the differences between the time series in the lower dimensional space. This is because our method explicitly accounts for the sequential nature of time-series; learning is performed in the space of readout weight representations and is guided by an objective function that quantifies the reconstruction error in a principled manner. Of course, the perfectly capable t-SNE is used here as a mere candidate from the class of algorithms designed to visualise vectorial data in order to demonstrate this issue. Moreover, we demonstrate that our method, by its very nature, is capable of projecting also unseen hold-out data. Future work will focus on processing large datasets of astronomical light curves.

References

[1] J. W. Richards, D. L. Starr, N. R. Butler, J. S. Bloom, J. M. Brewer, A. Crellin-Quick, J. Higgins, R. Kennedy, and M. Rischard. On machine-learned classification of variable stars with sparse and noisy time-series data. *The Astrophysical Journal*, 733(1):10, 2011.

[2] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *NIPS*, pages 487–493. The MIT Press, 1998.

[3] T. Jebara, R. Kondor, and A. Howard. Probability product kernels. *Journal of Machine Learning Research*, 5:819–844, 2004.

[4] H. Chen, F. Tang, P. Tino, and X. Yao. Model-based kernel for efficient time series analysis. In *KDD*, pages 392–400, 2013.

[5] H. Jaeger. The ''echo state'' approach to analysing and training recurrent neural networks. Technical report, German National Research Center for Information Technology, 2001.

[6] M. A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AICHE Journal*, 37:233–243, 1991.

[7] A. Rodan and P. Tino. Minimum complexity echo state network. *IEEE Transactions on Neural Networks*, 22(1):131–144, 2011.

[8] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[9] T. Gneiting and M. Schlather. Stochastic models that separate fractal dimension and the hurst effect. *SIAM Review*, 46(2):269–282, 2004.

[10] K. P. Harikrishnan, R. Misra, and G. Ambika. Nonlinear time series analysis of the light curves from the black hole system grs1915+105. *Research in Astronomy and Astrophysics*, 11(1), 2011.

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY
MNRAS **451,** 3385–3392 (2015) doi:10.1093/mnras/stv1181

Featureless classification of light curves

S. D. Kügler,* N. Gianniotis and K. L. Polsterer*
Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg, Germany

Accepted 2015 May 20. Received 2015 May 20; in original form 2015 April 15

ABSTRACT

In the era of rapidly increasing amounts of time series data, classification of variable objects has become the main objective of time-domain astronomy. Classification of irregularly sampled time series is particularly difficult because the data cannot be represented naturally as a vector which can be directly fed into a classifier. In the literature, various statistical features serve as vector representations. In this work, we represent time series by a density model. The density model captures all the information available, including measurement errors. Hence, we view this model as a generalization to the static features which directly can be derived, e.g. as moments from the density. Similarity between each pair of time series is quantified by the distance between their respective models. Classification is performed on the obtained distance matrix. In the numerical experiments, we use data from the OGLE (Optical Gravitational Lensing Experiment) and ASAS (All Sky Automated Survey) surveys and demonstrate that the proposed representation performs up to par with the best currently used feature-based approaches. The density representation preserves all static information present in the observational data, in contrast to a less-complete description by features. The density representation is an upper boundary in terms of information made available to the classifier. Consequently, the predictive power of the proposed classification depends on the choice of similarity measure and classifier, only. Due to its principled nature, we advocate that this new approach of representing time series has potential in tasks beyond classification, e.g. unsupervised learning.

Key words: methods: data analysis – methods: statistical – techniques: photometric – astronomical data bases: miscellaneous.

1 INTRODUCTION

The variation of the brightness of an astronomical object over time (hereafter called light curve or time series) is an important way to obtain knowledge and constraint properties of the observed source. With the advent of large sky surveys such as the Large Synoptical Sky survey (LSST; Ivezi et al. 2014), the incoming data stream will be so immense that the applied methodology has to be reliable and fast at the same time. While the origin of variability can be very different, a huge fraction of the variable objects in the sky has a stellar origin. From those variable stars many show (quasi-) periodic behaviour, and originate from the instability stripe in the Hertzsprung–Russell diagram or are multistar systems where the origin of the variability is the mutual occultation. The main focus of this work will be on periodic sources, but in principle the presented methodology can also be used for non-periodic sources (see e.g. Donalek et al. 2013).

The classification performance of periodic sources is already fairly high provided that the period and the amplitude of the variation are determined correctly (Bailey & Leland 1899; Bailey 1902; Bono et al. 1997). But apart from the very soft boundaries between the classes, the quality of the period-finding algorithm depends on the type of variability itself (Graham et al. 2013) and thus a dependence between those two properties is encountered. In order to break this dependence, one can either rely on only (quasi-) static features[1] for the classification or estimate the period and derive classifications by analysing the phase-folded light curves (see e.g. Debosscher et al. 2007, and references herein). Richards et al. (2011) showed that the inclusion of static features yields an improved classification

* E-mail: dennis.kuegler@h-its.org (SDK); kai.polsterer@h-its.org (KLP)

[1] Throughout this paper, we will divide features derived by other authors in three categories: non-static – everything directly related to period finding and features derived from the periodogram; quasi-static – features that treat the data as function instead of a time series, e.g. slope, linear trend; static – all features that treat the measured fluxes only as an ensemble and thus the temporal information is discarded, e.g. median, standard deviation. A complete list of features used here is given in Table 1.

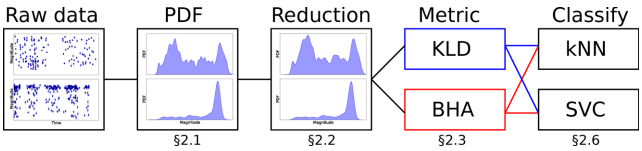


Figure 1. Schematic view of all steps from the raw data to the classification method.

performance and that the contribution of the static and non-static features to the accuracy is of the same order.

In this work, we introduce a novel representation of time series that aims to replace the static features. We represent each noisy data point by a Gaussian; the mean of the respective Gaussian is the measurement and the standard deviation is given by the measurement uncertainty (photometric error). Hence, every time series is represented by a mixture of Gaussians (MoG) that conserves all static information available in the data. We advocate that this is a simple and natural choice. In contrast to that, features can be seen as derivatives (such as moments) of this density model and therefore only describe certain properties of it. For instance, Lindsay & Basak (2000) show that moments are just able to describe the tails of a distribution but do not necessarily give a good description of the underlying distribution.

As a consequence, the proposed density-based representation presents an upper boundary to the static information content which can be made available to the classifier. The similarity of two densities is thereby judged using three widely used distance measures, the $L2$-norm, the Kullback–Leibler divergence (KLD) and the Bhattacharyya distance (BHA). These measures of similarity are then fed into two different classifiers. Finally, we compared the classification performance of the density- and feature-based approaches.

The aim of this work is to introduce an alternative and more general notion of similarity between light curves, which correctly takes into account measurement uncertainty. In the new representation, all static information contained in the observations are conserved in a more principled way and adjacently fed to the classifier. Consequently, we expect that this new representation provides a reference in terms of classification performance.

In Section 2, the new representation and its respective application to the classifier are described. After describing the used data in Section 3, the results of two different experiments are presented in Section 4. We conclude with a discussion of our approach in Section 5.

2 METHOD

In this section, the methodology is described. A sketch of the entire classification process is shown in Fig. 1. Each step is annotated with the respective subsection in the text; the FCLC software which includes all steps described in the following is available at http://ascl.net/1505.014.

2.1 Converting data points into densities

The key idea of our method is to convert the individual data points with their errors as a continuous density. We treat each data point as a normal distribution with a mean μ equal to the magnitude y and a width σ equal to the photometric error Δy of the respective

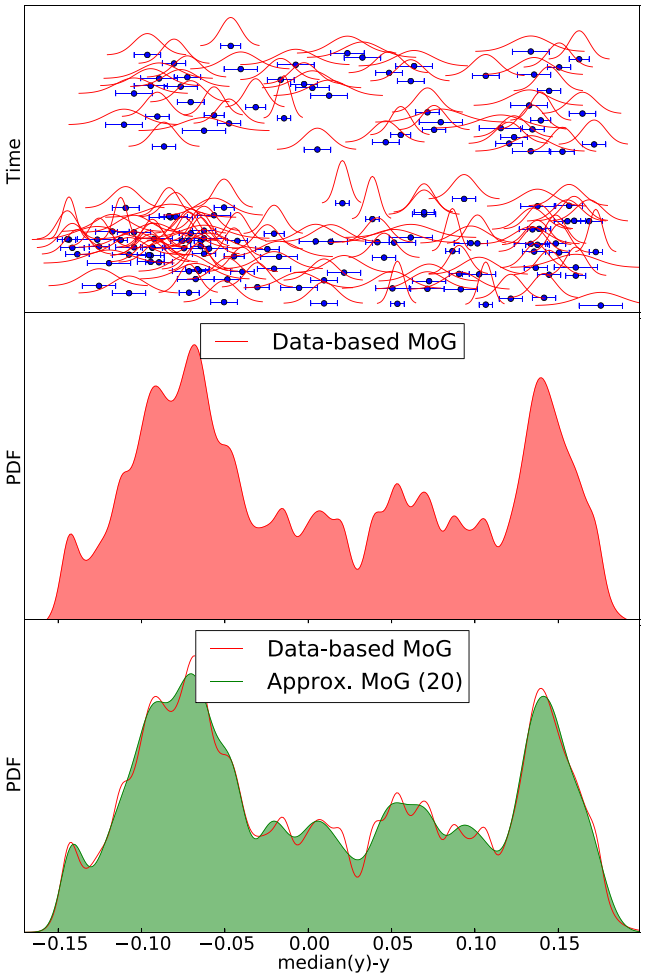


Figure 2. The principle of the conversion to densities, every point is described by a normal distribution which are then added up to a PDF.

measurement. This allows us to convert the discrete M number of observations into a continuous density by using

$$\mathrm{PDF}(x) = \frac{1}{M}\sum_{i=1}^{M}\mathcal{N}(x \mid y_i, \Delta y_i)\,, \tag{1}$$

where $\mathcal{N}(\mu, \sigma)$ is the normal distribution with expectation μ and width σ, which returns the probability of the occurrence for a given value x. Each light curve is, after subtracting the median, converted to such a probability density function (PDF); a visualization of this process is shown in Fig. 2. This idea was already mentioned in the work of Aherne, Thacker & Rockett (1998).

2.2 Parsimonious mixtures of Gaussians

An important step to make the computation of the distances computationally feasible is to reduce the number of Gaussians in the MoG. The look-up function for individual values of x scales linearly with the number of Gaussians in the mixture. The computation of the distance between every two densities scales directly with the number of Gaussians m in each density. Thus, the computational complexity for computing the distance matrix is of the order of $\mathcal{O}(n^2 m)$, where n is the total number of light curves to be classified. This computation gains a significant speedup by reducing the number of Gaussians; reducing the number of observations (typically $M = 300$) to an

MoG with $m = 20$ components yields an effective gain in speed of $(\frac{300}{20}) \approx 15$.

We tested several ways described in the literature to reduce the number of Gaussians effectively (Crouse et al. 2011). After experimenting with the different methods, we found that the method by Runnalls (2007) yielded the most satisfactory results. The basic idea is that two similar (in terms of the KLD, see below) Gaussians can be approximated by a single normal distribution. The dissimilarity between two normal distributions with amplitudes W_0; W_1 means μ_0; μ_1 and widths σ_0; σ_1 is thereby measured by

$$D = 0.5\omega \log\left(\tilde{\omega}_0 \frac{\sigma_0^{2\tilde{\omega}_1}}{\sigma_1^{2\tilde{\omega}_1}} + \tilde{\omega}_1 \frac{\sigma_1^{2\tilde{\omega}_0}}{\sigma_0^{2\tilde{\omega}_0}} + \tilde{\omega}_0\tilde{\omega}_1 \frac{(\mu_1-\mu_0)^2}{\sigma_0^{2\tilde{\omega}_0}\sigma_1^{2\tilde{\omega}_1}}\right) \quad (2)$$

with $\omega = \sigma_0 + \sigma_1$; $\tilde{\omega}_0 = \frac{\sigma_0}{\omega}$; $\tilde{\omega}_1 = \frac{\sigma_1}{\omega}$. The pair of normal distributions with the closest distance D is then merged into a new single Gaussian with weight $W_{01} = W_0 + W_1$, expectation $\mu_{01} = \frac{W_0}{W}\mu_0 + \frac{W_1}{W}\mu_1$, and variance $\sigma_{01}^2 = \frac{W_0}{W}\sigma_0^2 + \frac{W_1}{W}\sigma_1^2 + \frac{W_0 W_1}{W^2}(\mu_0 - \mu_1)^2$. The search and replacement is then performed iteratively until the desired number of new components is reached. An example of a reduced MoG is shown in the bottom plot in Fig. 2. Apart from the decreased computational complexity, the reduction in number of components used in the MoG has yet another very interesting side effect. Due to the loss of information, the new PDF is always just a smoothed version of the density based on the real data. As the data are irregularly sampled, this smoothing is effectively a better representation of the true underlying density. Obviously, the number of Gaussians to be used is a parameter which has to be optimized. Here, it will be optimized by maximizing the classification accuracy for a given data set and classifier.

Another aspect to mention is the conservation of outliers. Since iteratively only the most similar Gaussians are merged into a single one, the presence and probability of outliers will remain unchanged throughout this procedure.

2.3 Similarity of probability densities

After converting all light curves to PDF, we apply different measures of similarity between two given probability densities $P(x)$, $Q(x)$. As light curves differ in apparent magnitude, we subtract the median magnitude in order to align the densities of different objects.

2.3.1 *L2-norm*

The most obvious choice for comparing two densities is the *L2*-norm, defined as

$$L2\,(P(x),\,Q(x)) = \int (P(x) - Q(x))^2 \,\mathrm{d}x. \quad (3)$$

While the *L2*-norm is a very robust and reliable measure of the similarity, it is not very sensitive to faint tails as differences in the main components are penalized more heavily. But, as stated in the Introduction, the tails contain the vast majority of information of a density. Hence, we do not expect the *L2*-norm to be a good distance measure for our classification problem.

2.3.2 *Bhattacharyya distance*

The Bhattacharyya distance (BHA), defined as

$$\mathrm{BHA}\,(P(x),\,Q(x)) = -\log \int \sqrt{P(x)Q(x)}\,\mathrm{d}x, \quad (4)$$

is a generalization of the Mahalanobis distance which, in contrast to the latter one, takes into account the difference in shape. The BHA has been used in classification problems before (see e.g. Aherne et al. 1998) and thus seems a very good choice for our method.

2.3.3 *Symmetrized Kullback–Leibler divergence*

The Kullback–Leibler divergence (KLD),

$$\mathrm{KLD}\,(P(x),\,Q(x)) = \int P(x) \log\left(\frac{P(x)}{Q(x)}\right)\mathrm{d}x, \quad (5)$$

is a measure of similarity of two probability densities in information theory. It consists of two terms, one being the entropy (information content) of $P(x)$ and a term which is the expectation of $\log(Q(x))$ with respect to $P(x)$. The second term is the log-likelihood that the observed density $Q(x)$ was drawn from the model density $P(x)$. The KLD is capable of describing also difference between densities in faint tails. The KLD itself cannot be treated as a distance directly since – even though it returns zero for identical densities – it is not symmetric. We circumvent this problem by simply symmetrising the KLD and thus we finally compute

$$\mathrm{KLD}_{\mathrm{sym}}\,(P(x),\,Q(x)) = \mathrm{KLD}\,(P,\,Q) + \mathrm{KLD}\,(Q,\,P)\,. \quad (6)$$

2.4 Computation of distances

The KLD and the BHA cannot be computed analytically for two MoG and thus must be approximated by performing the integration. Even though analytical approximations exist for the KLD (see Durrieu, Thiran & Kelly 2012, and references herein), we encountered numerous difficulties when using them in practice, e.g. as non-positive distances. For this reason, we decided to perform the integration for all distances numerically. For our one-dimensional case, we found the following numerical integration to be sufficient:

$$\int_{-\infty}^{\infty} F(x)\,\mathrm{d}x \approx \Delta\,(F\,(x_0) + F\,(x_0+\Delta) + \cdots + F\,(x_1))\,. \quad (7)$$

The integration above is performed from $-\infty$ to $+\infty$. Here the integral is numerically approximated and therefore a finite range must be defined. The lower and the upper boundary are chosen very generously by integrating from $x_0 = \mu(i) - 5\sigma(i)$, with $i = \mathrm{argmin}_{i\in\mathrm{MoG}}\mu(i)$ to $x_1 = \mu(i) + 5\sigma(i)$, with $i = \mathrm{argmax}_{i\in\mathrm{MoG}}\mu(i)$. In order to retain the same precision for all integrals, we chose the integration width Δ for all integrations to be the same. To be on the safe side, we set $\Delta = 0.001$ but when experimenting with this width, it turned out that $\Delta = 0.005$ is sufficiently small to minimize computation time (scales with Δ^{-2}) without any loss in accuracy. Obviously, a good estimate for Δ is given by taking a fraction of the typical standard deviation in the MoG, as then the integration resolution is below the typical scale width of the density. To prevent the integration from encountering ill-defined (that is, negative values in the square root or log) values we add a small constant to each of the densities which does not yield any measurable impact on the final classification.

In Fig. 3, the impact of the injection of a single outlier on the *L2*, BHA, and KLD with respect to its injection position (x-axis) is shown. Therefore, a single measurement value with a typical photometric error is inserted into the distribution from Fig. 2 and the distance to the undistorted distribution is computed, respectively. It is evident, that the KLD reacts way more heavily to a single outlier, which will eventually also limits its use for classification task, as shown in the results. Note that in principle, the KLD distance would

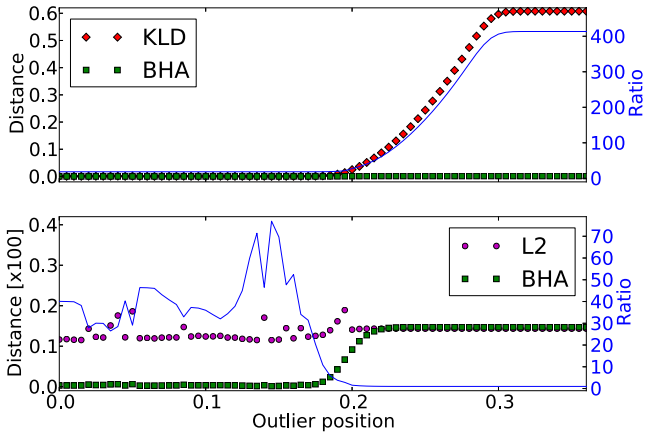


Figure 3. The effect of an outlier added to the distribution in Fig. 2 is shown. The *x*-axis gives the magnitude of the injected point with respect to the median, the left *y*-axis are the KLD/BHA distance in the upper plot, and the *L2*/BHA distance in the lower one. The right *y*-axis denotes the ratio of the respective distance measures.

Table 1. Computed features from observations and from the density model with respective formulas of an example light curve.

Feature	Moment	Data feature	Model feature
Amplitude	$A = 0.5 \cdot x_{0.95,\,0.05}$	0.150	0.141
Beyond1Std	$1 - \int_{\sigma_1-\sigma_2}^{\sigma_1+\sigma_2} P(x)\mathrm{d}x$	0.446	0.435
FPRMid20[a]	$x_{0.60,\,0.40}/x_{0.95,\,0.05}$	0.325	0.309
FPRMid35[a]	$x_{0.675,\,0.325}/x_{0.95,\,0.05}$	0.479	0.468
FPRMid50[a]	$x_{0.75.0.25}/x_{0.95,\,0.05}$	0.625	0.631
FPRMid65[a]	$x_{0.825,\,0.175}/x_{0.95,\,0.05}$	0.804	0.789
FPRMid80[a]	$x_{0.90,\,0.10}/x_{0.95,\,0.05}$	0.904	0.899
Skew	σ_3/σ_2^3	−0.185	−0.182
SmallKurtosis	$\sigma_4/\sigma_2^4 - 3$	−1.365	−1.361
MAD	x_{MAD}	0.083	0.083
MedianBRP	$\int_{x_{0.5}-A/5}^{x_{0.5}+A/5} P(x)\mathrm{d}x$	0.142	0.126
PercentAmpl.[b]	incl. median of LC	0.013	–
PDFP[b]	incl. median of LC	0.021	–
StetsonK[b]	incl. # observations	0.894	–

Notes. [a]FPR: FluxPercentileRatio, $x_{f,\,g} = x_f - x_g$
[b]Features without equivalent model description.

diverge to infinity; the plateau is just encountered due to the added small constant mentioned above.

2.5 Relation to features

Our density representation directly relates to the features[2] used in Richards et al. (2011); a detailed definition of all the used features is given in the Appendix A. As shown in Table 1, we can recover all but three features directly from the density, except for StetsonK, PercentAmplitude, PercentDifferenceFluxPercentile (PDFP). StetsonK contains the discrete number of observations as one of its input parameters, the latter two the absolute median value of the magnitudes. No equivalent measures for those exist for our median-subtracted and normalized densities.

[2] To compute the features we use the PYTHON package provided at http://isadoranun.github.io/tsfeat/FeaturesDocumentation.html, also used in Nun et al. (2014).

To explain all the other features we use the common notion of moments of a density

$$\sigma_n = \int_{-\infty}^{+\infty} x^n P(x)\mathrm{d}x \tag{8}$$

with $\sigma_0 = 1$, σ_1 being the mean, σ_2 the standard deviation and so on. Another frequently used integral is the percentile, x_f, where the density contains a certain fraction f, defined by

$$x_f : \int_{-\infty}^{x_f} P(x)\mathrm{d}x = f\,. \tag{9}$$

Additionally, the median absolute deviation (MAD) is defined as

$$x_{\mathrm{MAD}} : \int_{0}^{x_{\mathrm{MAD}}} P(x - x_{0.5}) + P(x_{0.5} - x)\mathrm{d}x = 0.5, \tag{10}$$

where $x_{0.5}$ is the median. One can see that most features can be expressed in terms of our PDF and thus the computed density contains most of the information encoded in the features.

2.6 Classification

In this subsection, we are describing the functionality and use of the different classifiers applied in this work. The first two of those classifiers depend actually on a distance matrix which is the direct outcome of our distance measure. For the features, the distance matrix is created by computing the Euclidean distance

$$D\,(v,\,w) \equiv \sum_{n=1}^{N_{\mathrm{feat}}} \sqrt{(v_n - w_n)^2} \tag{11}$$

between two feature vectors v, w. For the interested reader, more details on the used classifiers can be found in Hastie, Tibshirani & Friedman (2009). We use the implementations provided in the PYTHON package SCIKIT-LEARN.[3] To exclude effects originating from the preprocessing of the features, we also classified the light curves with a min–max normalized version of the features. In the following, the applied classifiers k nearest neighbours (kNN) and the support vector machine (SVM) are explained in more detail.

2.6.1 *k nearest neighbours*

Once the distances between all light curves are computed, we can sort the matrix for each candidate light curve and look at the types of the closest reference light curves. The only free parameter is k, the number of neighbours chosen per test light curve. Another degree of freedom can be introduced by weighting the distances to the neighbours, e.g. decaying distance. In practice, we obtained no significant gain and thus we use a classical majority vote. If the number of objects of a certain class is equal for two (or more) different classes, a random class out of those is assigned.

2.6.2 *Support vector machine*

A slightly better performance in classification can be reached if the distance matrix is used as the kernel of an SVM. In this work, we use the radial basis function (RBF) kernel which reads

$$K_{ij} = \exp\left(-\frac{1}{\delta^2} D_{ij}\right) \tag{12}$$

[3] http://scikit-learn.org/

Table 2. Overview of classifier parameters to be optimized.

Classifier	Parameter	Range
*k*NN	No. of neighbours k	1, 2, ..., 30
	Weights	Uniform (fixed)
ν-SVM	Softening parameter ν	0.01, ..., 1.00 (adaptive)
	Kernel width δ	0.01, ..., 100 (adaptive)
	Kernel	RBF (fixed)
	Kernel degree	3 (fixed)
RF	Number of trees T	100, 200, ..., 1000
	Split algorithm	gini (fixed)
	Max. number of features	All (fixed)

with D being the distance matrix and δ the bandwidth. As a consequence, low distances will have a kernel value close to unity and distances significantly larger than δ will be close to zero. We take the ν-SVM as the kernel classifier. Two parameters have to be tuned in a ν-SVM, namely the kernel width δ and the width of the soft margin ν. The soft margin controls the fraction of misclassifications in the training of the classifier.

2.6.3 *Random forest*

Given the success in Richards et al. (2011), we use the random forest (RF) as a candidate classifier as well. RF extends the concept of a single decision tree by using an ensemble of randomized decision trees. Unfortunately, by its very nature, this classification method can only be used on features. At each node of a tree, the features are split such that the information content (entropy) is maximized at each decision. The dominant free parameter in an RF is the number of decision trees, which is the only one considered in this work.

2.7 Performance and optimization

Each of the classifiers presented has several free parameters to be optimized but we stick for all the methods with the most important ones. For the *k*NN comparison this parameter is the number of investigated k nearest objects, the ν-SVM classifier has the tunable softening parameter ν, and the kernel width δ and eventually the RF can be built up of T number of trees. While other parameters (e.g. tree depth in RF) might have an impact on the classification quality, it is not the aim of this work to investigate this possible gain with the choice of these parameters. Also the process of feature selection is skipped and throughout this work always all features defined in Table 1 are used. All the parameters are evaluated for each classifier and data set independently on a fixed grid and the respective value with the highest accuracy is eventually chosen. A summary over the tuned parameters and their respective search ranges, as well as all parameters that have not been optimized, are shown in Table 2.

We judge the performance of a classifier by computing the accuracy defined as the mean fraction of correctly classified targets over a 10-fold cross-validation; the uncertainty in accuracy is given by the standard deviation. In addition, we compute the confusion matrix of the best classifiers to investigate possible caveats in the presence of multiple and unbalanced classes.

3 DATA

We conduct experiments with the different representations and classifiers on two data sets. This has the advantage that we have two independent measures for the predictive power of our method. In the first experiment, three classes are to be separated; in the second, a more complex seven class classification is performed. In fact, in the former data set the classes are defined more broadly (e.g. no distinction between different binary classes) and thus it is expected that the classification accuracy will be higher than in the latter case. It is the aim of this experiment to show, that our classification algorithm can perform comparably well to state-of-the-art classifiers for very broad and detailed classification tasks alike.

Table 3. Types of variables, number of entities and average number of observations.

Survey	VarType	Entities	⟨No. of obs⟩
OGLE	Cepheids	3567	225
	Eclipsing binaries	3929	330
	RR Lyrae	1431	323
ASAS	Mira	2833	342
	ED	2292	570
	RR Lyrae AB	1345	412
	EC	2765	524
	ESD	893	547
	DSCT	566	492
	DCEP-FU	660	561

3.1 OGLE

The Optical Gravitational Lensing Experiment (OGLE; Udalski et al. 2008) is a survey originally dedicated to the search for microlensing events and dark matter. Therefore, stars of the Magellanic clouds and the Galactic bulge were monitored for the unique traces of microlensing events. Consequently, millions of stars have been monitored, delivering a rich data base of variable stars. In our work, we use the data set used in Wang, Khardon & Protopapas (2012)[4] where some RR Lyrae, eclipsing binaries and Cepheids in the Magellanic clouds were extracted from the OGLE-II survey. The objects selected were known to be periodic before and thus their period was known as well. In the publication, the determination of the period is the main goal, but the data base presents a good test bed for classification as well, since a correctly determined period favours also good classification results and thus the classification is very reliable. The total number of objects is listed in Table 3. Some of the files contain lines with invalid entries, that is a few lines with a measurement error of zero, which have been removed.

3.2 ASAS

The All Sky Automated Survey (ASAS; Pojmanski 1997) is performed with telescopes located on Hawaii and in Las Campanas and is led by the Warsaw University in Poland. The sky is observed in the I and V bands with an initial limit of 13 mag (later extended to 14 mag). In 2005, the ASAS catalogue of variable stars (ACVS; Pojmanski, Pilecki & Szczygiel 2005) was published which is the starting point for our experiment. From the ACVS, we extracted all objects with a unique classification which is not miscellaneous. Subsequently, we removed all light curves having less than 50 observations and all classes with less than 500 members. A summary of the classes used can be found in Table 3. For the classification we used the magnitude 'Mag_2' (which corresponds to a 4 pixel aperture) which is a reasonably good measure of the brightness for

[4] www.cs.tufts.edu/research/ml/index.php?op=data_software

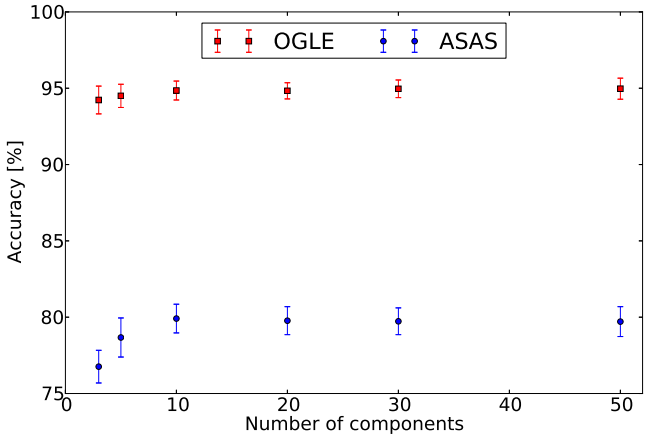


Figure 4. Classification accuracy versus the number of Gaussian components using the *k*NN classifier on both data sets. Accuracy is largely insensitive to the exact choice of the number of Gaussian components.

fairly bright and faint stars. Due to the extension to the faint end, the classes given could inherent some false classifications itself, especially since also subclasses (e.g. detached and contact binaries) are annotated and hence, it is expected, the class assignment in the given catalogue is not as reliable as in the OGLE case.

4 RESULTS

As stated in the methodology, the impact of the reduction of the number of Gaussians has to be quantified. In Fig. 4, we show empirically that the impact on the final accuracy is only marginal, as long as the number of components exceeds 10. For all conducted experiments, we fix the number of Gaussians to 20.

In Tables 4 and 5, the results of the different experiments for the OGLE and ASAS data set are shown, respectively. Since the *L2*-norm performs, independently of the chosen classifier, always worse than the BHA and KLD metrics, we exclude it from the discussion in the following.

For the OGLE experiment, we see that each method (feature and density methods) performs comparably well within the typical deviation between the 10 cross-validation folds. It is worth noting, the RF, claimed to be the best classifier in Richards et al. (2011), does not perform any better than the other classifiers. It is further

Table 4. Results for the optimal classifiers for the three-class classification of OGLE data. The performance is the average fraction of correctly classified objects in a 10-fold cross-validation with the standard deviation of this performance being the error (all given in per cent).

	*k*NN	ν-SVM	RF
Features (raw)	95.09 ± 0.74 $k = 8$	96.86 ± 0.52 $\nu = 0.04, \delta = 0.31$	95.61 ± 0.82 $T = 500$
Features (norm.)	95.51 ± 0.81 $k = 10$	96.88 ± 0.67 $\nu = 0.06, \delta = 0.08$	95.59 ± 0.83 $T = 500$
L2	93.44 ± 0.88 $k = 3$	95.92 ± 0.68 $\nu = 0.06, \delta = 0.69$	– –
KLD	95.14 ± 0.70 $k = 5$	95.51 ± 0.94 $\nu = 0.14, \delta = 0.33$	– –
BHA	94.84 ± 0.83 $k = 7$	96.01 ± 0.71 $\nu = 0.08, \delta = 0.14$	– –

Table 5. Results for the optimal classifiers for the seven class classification of ASAS data. The performance is the average fraction of correctly classified objects in a 10-fold cross-validation with the standard deviation of this performance being the error (all given in per cent).

	*k*NN	ν-SVM	RF
Features (raw)	74.22 ± 1.24 $k = 11$	78.02 ± 0.68 $\nu = 0.19, \delta = 0.53$	79.98 ± 1.16 $T = 400$
Features (norm.)	77.60 ± 0.76 $k = 17$	80.47 ± 1.21 $\nu = 0.17, \delta = 0.10$	79.99 ± 1.55 $T = 400$
L2	79.57 ± 0.80 $k = 19$	82.08 ± 0.89 $\nu = 0.01, \delta = 0.56$	– –
KLD	78.96 ± 1.87 $k = 23$	75.56 ± 0.94 $\nu = 0.26, \delta = 0.34$	– –
BHA	79.73 ± 0.83 $k = 29$	81.11 ± 0.90 $\nu = 0.20, \delta = 0.14$	– –

interesting to see that the feature-SVM is performing slightly better than the SVM based on the density representation. As mentioned in Section 2, three features exist which cannot be described by the density-based approach. When removing those respective features from the feature list, the accuracy of both feature-SVMs drops by one per cent, indicating that the difference in accuracy does originate from those. The strength of the variation with respect to the median observed brightness appears to bear some information about the type of variability. We elaborate further on this issue in the discussion section. That the impact of those median-based features is anyway not too high is supported by the results of the seven class ASAS classification. It becomes apparent that the more generic definition of the density enhances the accuracy in contrast to all the feature-based classifiers. The confusion matrices of the best classifiers from the density and feature based classification are shown in Figs 5 and 6. It can be seen that classes with more members achieve a higher accuracy which is expected due to the higher number of training objects. Otherwise, no significant biases in any direction between the two different classification approaches can be detected. While the gain in accuracy is again only marginal, it can be shown that the same quality is only reached if the three features, not describable by the densities, are included. Else the classification rates of the feature-based SVMs drop again by 1 per cent. Apart from this, it can be observed that the classification quality of the density-based

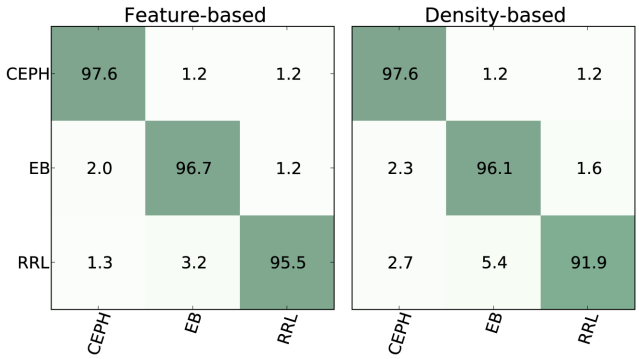


Figure 5. The accuracies (given in per cent) of the feature based (left) and density (BHA-metric) based ν-SVM classifier are shown for the OGLE data set. The *x*-axis shows the labels according to the classifiers, the *y*-axis the given ones; the colour scale stands for the respective accuracy; from zero (white) to hundred (green) per cent.

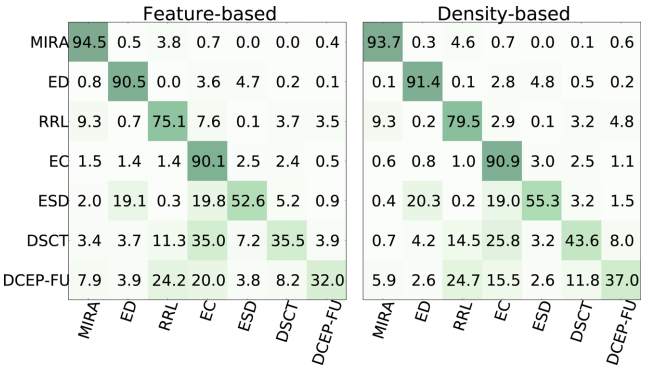


Figure 6. The accuracies (given in per cent) of the feature based (left) and density ($L2$-metric) based ν-SVM classifier are shown for the ASAS data set. The x-axis shows the labels according to the classifiers, the y-axis the given ones; the colour scale stands for the respective accuracy; from zero (white) to hundred (green) per cent.

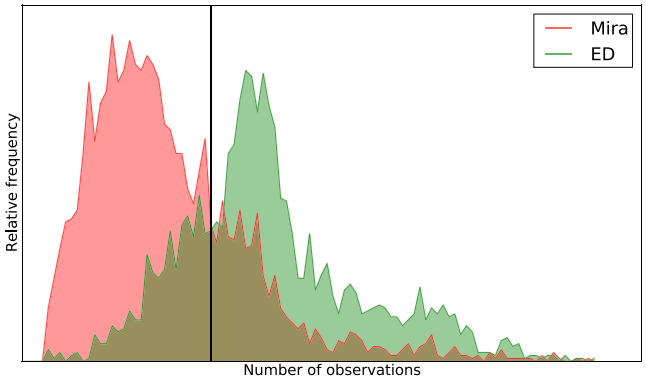


Figure 7. A histogram over the number of observations for the Mira and detached binaries classes in the ASAS survey are shown. The number of observations clearly correlates with the class label: if a bisectional line is introduced at 427 observations, a classification rate of 75.1 per cent can be reached.

classifiers depends quite strongly on the choice of the distance metric. The KLD does perform in three of the four experiments worse than the BHA which supports the statement that the BHA distance is a good distance measure for classification tasks. On the other hand, one should realize that the choice of the metric, that is the distance between two given feature vectors, is in principle, also a free methodological factor in the classification problem. Apart from the standard Euclidean distance and the Mahalanobis distance no other measures have been investigated in the literature.

5 DISCUSSION

In this work, we present a generalization of static features for the classification of time series. In contrast to previous work, we do not rely on describing static densities with a set of features but use the densities themselves to measure the similarity between two light curves. By doing so, we can reduce the number of degrees of freedom in the methodology from four (pre-processing, feature selection ,choice of metric, choice of classifier) to two (choice of metric and choice of classifier). This allows us to skip the step of feature selection. The proposed approach follows first principles by simply assuming a model for representing the data; once a metric is chosen, classification in a kernel setting follows naturally. The strong point of the newly proposed representation is the fact that it captures all the information present in the data (including measurement errors) and makes it available to the classifier.

As highlighted in the results, the choice of the metric used in the density representation plays an important role. A priori, we are not aware of any natural choice of a metric. We have shown in our experiments that the BHA and $L2$ distance are performing very well in terms of accuracy. In principle, other (or combinations of) metrics might exist that are more suited for a given classification problem.

Our approach presents a different way of performing classification. Therefore, it provides independent evidence that the widely used features are indeed well chosen for the classification problems considered so far. However, it is unclear how well the chosen features generalize to other classification problems. On the other hand, the density representation is formulated generically and encodes all information available in the data. Additionally, the proposed method naturally encodes also uncertainty in the measurements, which is not taken into account in the feature-based approaches so far. As a consequence, it is now possible to learn a classification on data of one survey that contains small (large) measurement uncertainty, and predict on data of another survey with large (small) photometric error. While this is problematic for feature-based approaches, it is automatically taken care of in the density representation.

We have shown that the feature- and density-based approaches perform comparably well in terms of accuracy for the given data sets. As aforementioned, there are three features that cannot be derived from the density representation which appear to increase the classification accuracy. In particular, the StetsonK value depends directly on the number of observations in a light curve, and for this reason it cannot be derived from our representation. It is questionable why the number of observations should be a defining property of a class. The only reason why it contributes to the performance is because certain classes are apparently observed more often than other ones (see Table 3) and not because it is an inherent physical property. In Fig. 7, we show that it is possible to classify ASAS light curves into Mira and detached binaries with a 75 per cent accuracy solely using the number of observations that happened to be recorded. The brightness of stars that vary over a wide range of magnitudes, such as MIRA, will frequently drop below the survey-specific detection limit. Hence, faint observations will not be recorded in the data base. This raises the following problem: absent recordings are ambiguous because it is not evident whether the source was too faint to be detected or simply not observed. As a consequence, the number of observations, and thus StetsonK, hints to the variability type of a star within one survey. However, this feature is survey dependent as surveys differ in data base structure (e.g. some give upper limits) and detection limits, and thus does not generalize. If non-detections are not treated accordingly, the definition and use of the StetsonK value can cause dramatic bias on the classification, especially when knowledge is transferred between different surveys, as done, e.g. in Blomme et al. (2010). Similarly, the PercentAmplitude and the PDFP directly depend on the apparent magnitude of the respective object, which is also not an inherent property of a class. Conclusively, the only reason why these three features contribute to the accuracy is because of the presence of a (or several) bias in the observations and not because they capture physical characteristics of the data. We do not state that the features in question are useless for classification (indeed they increase the accuracy), but argue that they do not generalize and are therefore not useful for knowledge transfer between surveys with different observational bias. It should be considered, to redefine these

features accordingly, such that they do not rely on the observation strategy of a survey.

In summary, the proposed method (a) introduces a more general notion of distance between light curves in contrast to static features, (b) naturally incorporates measurement errors, (c) performs equally well as state-of-the-art feature-based classifications, and (d) yields an independent measurement of the accuracy as compared to feature-based classification.

As a future prospect, the density-based representation could be useful in unsupervised settings where the notion of distance is more critical in the absence of labels which are the driving force in a classification task. Feature sets that have been optimized for classification do not necessarily provide a good similarity measure. In subsequent work, we will investigate whether the proposed notion of distance naturally distinguishes between the different variability types. Additionally, we advocate that besides static features also temporal information should be incorporated in a similar vein. However, the design of such a time-dependent representation remains an open question.

ACKNOWLEDGEMENTS

SDK would like to thank the Klaus Tschira Foundation for their financial support. The authors would like to thank the anonymous referee for his interest and his very helpful suggestions.

REFERENCES

Aherne F. J., Thacker N. A., Rockett P. I., 1998, Kybernetika, 34, 363
Bailey S. I., 1902, Ann Harv. Coll. Obs., 38, 1
Bailey S. I., Leland E. F., 1899, ApJ, 10, 255
Blomme J. et al., 2010, ApJ, 713, L204
Bono G., Caputo F., Castellani V., Marconi M., 1997, A&AS, 121, 327
Crouse D. F., Willett P., Krishna P., Svensson L., 2011, A Look at Gaussian Mixture Reduction Algorithms, 14th International Conference on Information Fusion, Chicago, Illinois
Debosscher J., Sarro L. M., Aerts C., Cuypers J., Vandenbussche B., Garrido R., Solano E., 2007, A&A, 475, 1159
Donalek C. et al., 2013, preprint (arXiv:1310.1976)
Durrieu J. L., Thiran J. P., Kelly F., 2012, Lower and upper bounds for approximation of the Kullback-Leibler divergence between Gaussian Mixture Models, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
Graham M. J., Drake A. J., Djorgovski S. G., Mahabal A. A., Donalek C., Duan V., Maker A., 2013, MNRAS, 434, 3423
Hastie T., Tibshirani R., Friedman J., 2009, The Elements of Statistical Learning, 2nd edn. Springer, Berlin
Ivezi Z. et al., 2014, preprint (arXiv:0805.2366)
Lindsay B. G., Basak P., 2000, Am. Stat., 54, 248
Nun I., Pichara K., Protopapas P., Kim D.-W., 2014, ApJ, 793, 23
Pojmanski G., 1997, Acta Astron., 47, 467
Pojmanski G., Pilecki B., Szczygiel D., 2005, Acta Astron., 55, 275
Richards J. W. et al., 2011, ApJ, 733, 10
Runnalls A. R., 2007, Kullback-Leibler Approach to Gaussian Mixture Reduction, IEEE Transactions on Aerospace and Electronic Systems, p. 989
Stetson P. B., 1996, PASP, 108, 851
Udalski A. et al., 2008, Acta Astron., 58, 329
Wang Y., Khardon R., Protopapas P., 2012, ApJ, 756, 67

APPENDIX A: DETAILED DESCRIPTION OF FEATURES

In the following, we give a detailed description of the static features used in Richards et al. (2011). The computation of the software is done using the PYTHON FATS package, available under https://pypi.python.org/pypi/FATS. The error in the definition of the StetsonK value in older versions was corrected manually.

Amplitude: Absolute difference between the highest and the lowest magnitude.

Beyond1Std: Fraction of photometric points that lie beyond one standard deviation with respect to the (with photometric errors) weighted mean.

FluxPercentileRatio (FPR): Relative difference of flux percentiles with respect to the 95–5 percentile difference. The number after the FPR gives the width of the percentile, always centred on 50, e.g. $\mathrm{FPR20} = \frac{F_{60}-F_{40}}{F_{95}-F_5}$.

Skew: The skew of the distribution of magnitudes.

SmallKurtosis: Kurtosis of the magnitudes for small samples.

Median absolute deviation (MAD): Median deviation of the absolute deviation from the median.

Median buffer range percentage (MedianBRP): Fraction of data points lying within one-tenth of the amplitude around the median.

PercentAmplitude: Largest absolute difference from the median magnitude, divided by the median magnitude itself.

PercentDifferenceFluxPercentile (PDFP): The 95–5 flux percentile difference, divided by the median of the flux.

StetsonK: More robust measure of the kurtosis, as defined in Stetson (1996).

This paper has been typeset from a TEX/LATEX file prepared by the author.

Chapter 4

Discussion

The results of the individual publications were already discussed at an earlier point. In this chapter the usefulness and implications of these new approaches are discussed and the obtained scientific results are put in a broader context. The following considerations are discussed for each presented publication, respectively.

1. *Generalization.* The model generality refers to how flexible a model is. In the best case, the applied model is generic enough to be used on any given dataset without fine-tuning but is neither susceptible to over-fitting nor to over-simplification.

2. *Applicability.* The newly developed methodology should be eventually employed on large existing databases. Therefore, these approaches have to be feasible in terms of computational complexity (scaling) and reliability, e.g., in the sense that they do not depend on starting conditions too heavily.

3. *Advances compared to model-driven approaches.* The initial motivation to introduce a new methodology was to analyze databases without the deficits of standard model-driven approaches. It is therefore necessary to justify the need of new methodology and highlight the advantages and weaknesses of each approach, respectively.

4. *Insights gathered.* A new methodology is considered useful if it is validated with respect to existing methodology and is able to provide some additional benefit. It is argued why the revealed knowledge remained undetected with the former approaches and could be discovered by the presented ones.

5. *Future work.* Eventually, the presented approaches will be embedded in a broader context and possible limits and extensions are discussed. Some work under current consideration is also presented.

In summary, two central questioned will be answered in this section.

Was new methodology needed?
Can the presented approaches eventually replace existing ones?

4.1 Determination of spectroscopic redshifts

Generality The presented approach to extract possible multiple redshifts in spectra generalizes very well, as long as the training and testing set come from the same wavelength-regime and the spectra are of a comparable resolution. Since the data items are compared directly to each other, the number of parameters/methods that have to be chosen are only very few. The preprocessing has only a minor impact on the results, but changing the regression approach (here, k nearest neighbors) can have quite dramatic impact on the results. The algorithm can further be trained on any other provided dataset, e.g., coming from the infrared. Apart from estimating redshifts, other problems could be solved by modifying the preprocessing. The comparison of the continuum behavior could, for example, be used to estimate the strength of the Calcium break in spectra. Additionally, an initial clustering could be performed by using the same approach.

Applicability In practice, the methodology is easily applicable and due to the use of a non-parametric regression algorithm, the number of degrees of freedom for this algorithm is manageable. On the other hand, the computational expense of applying this methodology to a huge dataset is fairly high, the computational complexity scales with $\mathcal{O}\left(N^2 D\right)$. It should, however, be noted that the extracted feature vectors contain a considerable amount of entries that are close to zero. Performing dimensionality reduction beforehand, can limit the computational cost of the algorithm. This decrease of cost can thereby not only be attributed to the lower dimensionality of the representation (which changes the complexity linearly), but also due to a more efficient search structure in the comparison by using spatial tree structures (see, e.g., Bentley, 1975). Additionally, time can be saved by replacing the very robust and reliable leave-one-out cross-validation by a smaller training set. However, the loss in accuracy and completeness as a function of the size (and selection) of the training set have to be closely monitored as otherwise systematic (selection) effects may occur.

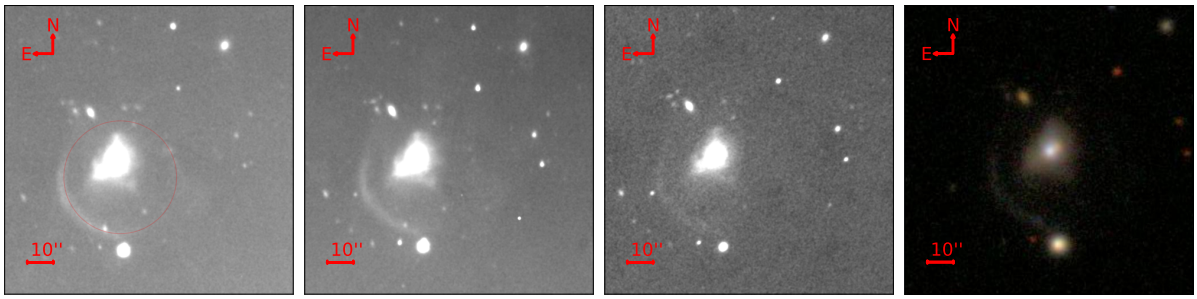
Advances The detection of multiple-redshift objects in the spectral database of SDSS can also be solved in a purely model-driven way. However, this introduces a lot of problems which limit the model-based approach. First of all, the simultaneous fitting of a template, spectral features and the redshift causes an unnecessary over-complication of the model. This is especially troublesome, if multiple (say r) redshifts are to be detected, as then the complexity of the model scales with

$$\mathcal{O} \sim \left(N_{\text{templates}} \cdot N_{\text{redshifts}}\right)^r > (50000)^r$$

where $N_{\text{templates}}$ is the number of templates and $N_{\text{redshifts}}$ is the number of tested redshifts. Apart from the complexity, an effective mechanism has to be chosen that can select the best-performing model. As shown in Subsection 2.4.1, this cannot be solely done on the basis of minimizing the difference between fit and data as then the model with the largest value of r is always favored. The possible solutions to that are regularization and/or splitting the data into training and validation set. Other approaches used for detecting multiple line system focus mainly on individual spectral regions where a local fit is employed to some pre-defined lines. These assumptions limit the use (and the generality) of these techniques dramatically as, e.g., in the case of an SMBHB also just one SMBH could be active while the absorption lines of the host galaxy indicate a relative movement of the active SMBH. However, it should be noted, that also the presented approach is not yet capable of describing an arbitrary number of superimposed redshift systems. The first experiment allowed only for a shift between emission and absorption, the second one between pre-defined emission and absorption regions. Apart from lines that might be missed in these selected regions, the use of the MAD for measuring the confidence of

a detected redshift system is still questionable. Especially, the low fraction of less than 60% of *good* redshifts indicates that more appropriate measures should be investigated. This statement is supported by a large fraction of *noisy* objects for which the redshift was anyway determined correctly. Conclusively, a purely model-driven approach can explain *expected* behavior much better, then the presented one, however, when it comes to extraordinary objects, which was the initial motivation, a data-driven approach might be of big help.

New insights The application of the newly introduced methodology revealed objects which have not been noted to be special before. All of the proposed objects have been flagged either with "SMALL_DELTA_CHI2" or "MANY_OUTLIERS" by the spectroscopic pipeline of SDSS. The pipeline denotes that something went wrong, however, the fraction of objects with either one of the flags is way too high to closely inspect the nature of all of them. The pipeline casts warnings quite generously which makes them fairly useless for the detection of spurious spectra. In summary, the existing pipeline, as well as other existing approaches, are not able to detect those outliers in an unbiased way, as the presented approach does. Therefore, it makes it an extremely powerful tool for investigating huge spectral (or generally high dimensional) databases. The three most prominent examples have been shown in the respective publication, and for one of them further investigation was carried out in order to confirm the existence of a giant arc. This investigation is illustrated in Figure 4.1 where observations in the sdss_g, sdss_i and Ks, obtained with MODS and LUCI (both instruments at the LBT) are shown. In all bands the arc is clearly visible and in the g-filter slight evidence can be seen, that the arc is actually part of a full circle surrounding the object (marked by a red circle). However, even with the generous exposures of 5 (13) min for MODS (LUCI) at a 8.4 m telescope only a fraction of the arc can be resolved. The radius of the circle is of the order of 20″ and it is interesting to see that its center does not coincide with the center of the very nearby galaxy. The object



(a) *sdss_g, 5 min* **(b)** *sdss_i, 5 min* **(c)** *Ks, 13 min* **(d)** *ugriz, 53.9 sec*

Figure 4.1: *Images of the lens candidate SDSS J120419.07-001855.93 observed at the LBT. Observations* ***(a)****,* ***(b)*** *performed with MODS;* ***(c)*** *with LUCI and* ***(d)*** *is a false-color image from the SDSS survey. In the first image, the potential ring/arc is highlighted.*

was also detected in Goto (2005) and was classified as an E+A galaxy (Dressler and Gunn, 1983) at redshift $z = 0.094$. In the SDSS spectrum very prominent Balmer absorption and no significant emission from star-formation is visible which supports this interpretation further. Even though, tidal tails are not uncommon for this galaxy type, it is interesting to see that the arc-like structure seems at no point connected to the central galaxy. While this does not rule out a tidal tail per se, it leaves at least some doubts about the origin of the segment of a circle. However, the large extend of this structure weakens the hypothesis of a gravitational lens since the number of lenses as extended as this is very limited so far (see, e.g., Muñoz *et al.*, 1998); a huge mass would be required to explain the observed ring. Further observations are needed to reveal the true origin of the observed structure, potentially, a spectroscopy would help to gain

further insight. However, this object is the best example for highlighting the advantages of the presented approach. In Fig. 4.2, one can see that, despite the blue end, the spectrum is nearly perfectly describable by a single blackbody radiator with some superimposed absorption lines. Since the SDSS pipeline is trying to fit spectral templates and redshifts simultaneously but does not allow for a large redshift for stellar templates at the same time, it was not able to describe that object accordingly. On the other hand, the newly presented method breaks the degeneracy between template selection and redshift estimation and can therefore more reliably estimate the redshift of this object.

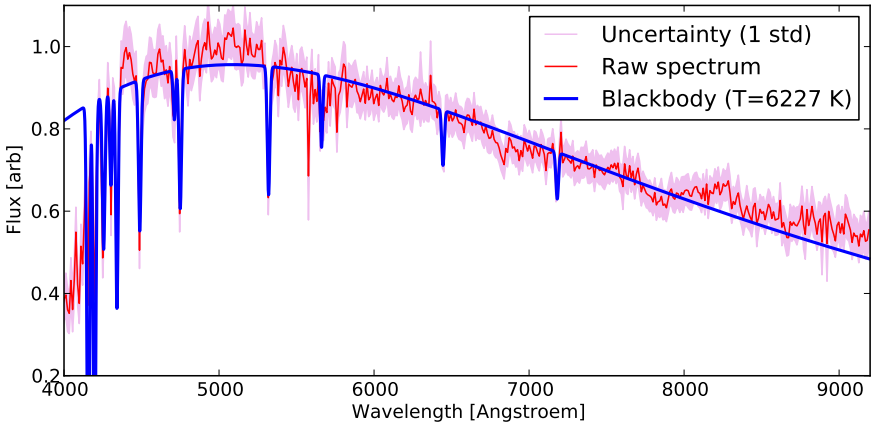


Figure 4.2: *SDSS Spectrum of the lens candidate SDSS J120419.07-001855.93. Planck's law, with superimposed absorption lines (see Tab. A.1) at redshift $z = 0.094$, is fitted to the data.*

Future work The k nearest neighbor regression is a useful alternative (or more of an extension) to the state-of-the-art models used for analyzing spectra. With its huge flexibility and ability to detect outliers, it can be used to check results from other analysis pipelines for correctness. Additionally, valuable knowledge hidden in the data can be revealed. So far, the algorithm still relies on a set of given training objects (supervised learning). With some modifications of the extracted model and the distance measure, this algorithm could also be transformed into a unsupervised database where the elements are sorted by the shift along the y-axis. In practice, this scenario is harder to realize as the distance has to be defined such that it is invariant against the amplitude of the spectral features and the continuum. The next steps of this model should be to test how the computational complexity can be reduced in an effective way and how the extracted sparse feature vectors could be turned into a lower dimensional structure. Finally, the application of the algorithm to the full SDSS database is desirable, as then a large number of peculiar objects can be identified and the performance of the presented algorithm against the existing approaches can be evaluated.

4.2 Visualization of time series

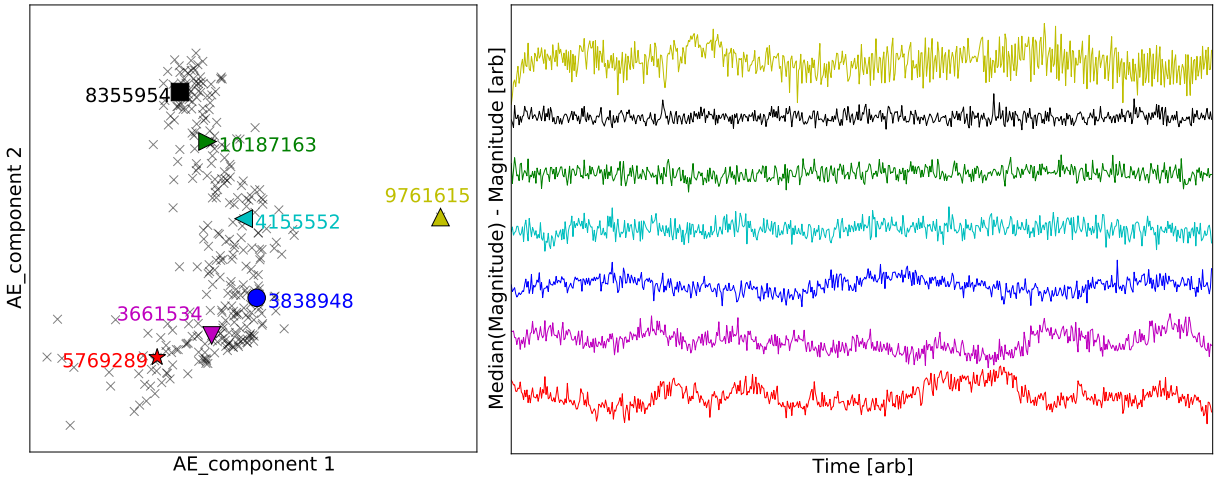
Generality The echo-state network (ESN) is a discrete time recurrent neural network and is the underlying model in the presented model-coupled autoencoder (AE). The ESN is able to capture the latent behavior that is causing the dynamical behavior of a time series. The presented method can be extended to visualize all kinds of data by solely replacing the ESN with a model that suites the presented science case. This highlights the broad applicability and generality of the proposed algorithm, which can be extended to symbolic sequences (e.g., text).

Applicability The application of the ESN-coupled AE on real data introduces some problems. First of all, the data have to be normalized in advance such that value range does not exceed the interval $[-1, 1]$. Secondly, the optimization of the AE is fairly costly. It includes the repetitive (expensive) use of the ESN, as eventually the reconstruction error should be measured directly on the data (time series, in this case). Additionally, in the presented case the AE has 2.040 trainable weights which have to be optimized. This optimization is not only costly, but convergence depends on initial random conditions. This is an undesired but inevitable behavior. To reduce the impact of the initial conditions, the optimization could be performed several times with different initial values, and eventually only the model with highest likelihood is kept. However, this would lead to another significant increase of computation time[14]. Despite the size of the ESN reservoir and the size of the hidden layer of the AE, no other initial parameters have to be set. Therefore, this methodology can be applied to any kind of unknown dataset without any further tuning.

Advances The direct application of dimensionality reduction algorithms on sequential data is not meaningful. This is because time series can be of variable length, should be treated in a shift-invariant way and eventually the sequential nature (neighboring data points are not independent) should be respected. All of those requirements are not met if the sequence is treated merely as a vector. In order to visualize sequential data, they have to be transformed into a vector representation, using an according model. This could be, for example, a physically motivated law such as the temporal behavior of an eclipsing binary. This would yield a much more meaningful representation than applying dimensionality reduction on interpolated, and phase-folded binary light-curves, as done in Matijevič *et al.* (2012). However, the choice of the model is dictated by the provided data and therefore drives the visualization. The model parameters returned by applying the model to the time series are then to be visualized. If the number of parameters is sufficiently low the parameter space can be inspected visually by creating the respective subplots. This visual inspection becomes already intractable for a simple binary model and therefore, the application of a dimensionality reduction algorithm is inevitable. Consequently, without the data-driven methodology a visualization of time series would not be possible. The huge advantage of this methodology is that it is capable of visualizing huge amounts of data in an unsupervised fashion and is therefore tailored to be applied on databases produced by Kepler or similar. The visualization algorithm in its current state is missing the capability of dealing with periodic behavior if the periodicity is longer than the memory of the ESN. ESNs with jumps or the inclusion of time-warping in the ESN are potential alternatives to incorporate periodic behavior quite naturally.

New insights Unfortunately, no other unsupervised methodology was used on non-periodic time series data so far and therefore, the performance of the presented approach cannot be judged. However, the ESN-coupled autoencoder provides an extremely powerful method to inspect huge database of sequential data in an explorative fashion. To highlight the importance and necessity of the presented approach, light curves as obtained from the Kepler survey are visualized in Figure 4.3a. From the catalog by Debosscher *et al.* (2011) sources with little evidence for periodicity ($Pf_1 > 0.5$), but considerable variability ($amp_{11} > 2e^{-4}$) were selected and only a subset of 303 smoothly variable objects has been kept. In the visualization a clear (arc-like) structure becomes visible with some overdensities around the black and the magenta markers. This highlights that also the non-periodic light curves show common behavior and can therefore be clustered using the presented approach. In order to validate the correctness of the

[14] On the other hand, multiple individual processes are straight-forward to parallelize. Thus, using as many cores as desired initial value settings would effectively not increase the computation time at all.



(a) *Visualization* **(b)** *Light curves of points marked in* **(a)**

Figure 4.3: *Visualization of non-periodic Kepler light curves. On the left side (a) the visualization using the ESN-coupled autoencoder is shown. The light curves of the marked objects are plotted in (b), the y-axis is scaled arbitrarily for each curve for better visibility.*

visualization, six light curves from distinct regions of the arc are plotted in Figure 4.3b. There, one can see that all objects located in the upper part of the visualization (green and black) are dominated either by random variability or consist of pure noise. The points further down in the representation exhibit some long-term variability (blue) and eventually the long-term behavior mixes with events on shorter time scales (red). It is also worth noting that the area between the blue and the magenta point is populated with light curves which exhibit long ($\sim 15\,d$) quasi-periodic behavior. The majority ($> 75\%$) of the visualized light curves have either been marked as *miscellaneous* or as *active stars* in the catalog of Debosscher *et al.* (2011). This shows the large potential in the application of visualization methods as it allows an unbiased view on the similarity of the given light curves. According to the NASA Exoplanet Archive[15], the outlier (KIC 9761615) on the very right of the plot is a potential exoplanet transit. Even though, the inspection of the first quarter of the light curve does not allow detailed conclusions it appears that this source is an active and strongly varying one rather than a good candidate for an exoplanet. In recent work (Kügler *et al.*, 2015)[16], the application of the proposed method on a larger dataset of Kepler lightcurves was performed. There, quite striking correlations between the variability behavior and the physical properties of the stars (such as temperature and surface gravity) were detected. It is pleasing to see that already known correlations between the variability behavior and physical properties were re-detected in the proposed visualization. Thereby no physical knowledge, such as periodicity, was imposed.

Future work The ESN-coupled AE is a unique tool to visualize sequential data in a fashion that preserve the similarity (dictated by the model) between data and not just the similarity between model parameters. The use of the ESN as the underlying model makes it especially useful for the visualizing regularly sampled astronomical time series. Even though, the number of tunable parameter is low, it is very flexible and can be applied to any given sequential data without much tuning. More efforts should be invested in decreasing the arbitrariness of the visualization introduced by the random initial state. Current work focuses on the inclusion of periodic behavior, as well as the use of physically motivated models.

[15] http://exoplanetarchive.ipac.caltech.edu
[16] http://arxiv.org/abs/1508.03482

4.3 Featureless classification

Generality The representation of the static part of an irregularly sampled light curves by a parsimonious mixture of Gaussian is a very natural one. The conversion to a probability density allows the use of the same classifier throughout different datasets. The major advantage of the approach is that observational biases are either considered (photometric errors) or are circumvented (sampling) such that the representation quality only depends on the number of observations and how independently they have been drawn. This opens also the possibility to perform an initial clustering on an unseen dataset as a generic distance can be computed which is not depending on the choice of handcrafted features. Therefore, the approach is very general. The down-side is that only parts of the available information are used because, so far, the dynamical information is discarded.

Applicability The conversion of the light curve into a parsimonious mixture of Gaussian is a computationally inexpensive task. However, the computation of the distance matrix is very costly, it scales with $\mathcal{O}\left(N^2\right)$. The distance between two mixtures of Gaussians can be approximated, however, testing different boundaries yielded very imprecise approximations, that were not sufficient for classification. Apart from the computational costs, the algorithm is easily applicable as it has only one free tunable parameter for the mixture of Gaussians and another one for the kNN (two for the SVM) classifier.

Advances Employing a purely model-driven approach for classification to irregularly sampled light curves is a very difficult task as mentioned before. On the other hand, data-driven approaches have been established for classification tasks for quite a long time. The use of purely data-driven methodology introduces certain issues that have to be considered. The creation and selection of features has been the central way of classifying light curves. However, in none of the presented publications the usefulness and reliability of the selected features has been evaluated properly. So far, the usefulness of features has been solely tested in terms of classification performance in in-sample tests. This is an extremely dangerous way of validating the performance because it mocks a correct treatment of survey-specific biases. The biases are the same for all observations within one survey. If the biases and uncertainties are not taken into account accordingly, the classifier itself will be biased instead. This limits the generality of the feature-based approach as the importance of the individual features is changing for each provided dataset. Even worse, the importance can even change for a single database if the selection of the training and testing set is not absolutely random. This problem can be only circumvented if a classification algorithm is employed which does not allow the introduction of a bias. In the presented experiments, it was shown that the classification performance decreases by 1% if biased features are excluded. While this seems a fairly low fraction it should be considered, that this refers to in-sample test. It is urgently required to investigate the impact of those biased features in a setting where the training and testing set have distinct biases. Besides the bias of the features, the impact of the uncertainty of the measurements could be investigated by learning on a faint (high photometric errors) subsample and predict on a bright (low error) one of the same survey. Including the uncertainties of the photometric measurements allows a comparable classification performance for noisy sources and additionally non-detections (that are upper limits) can be taken into account quite naturally, as opposed to the deterministic feature-based approach.

New insights The presented methodology enables the community to cross-check the results of their respective classification algorithms to a very well-defined standard classification algorithm. Newly provided features can be tested for their biases as in principle the classification

performance of the presented algorithm solely depends on the chosen metric (and less on the classifier). Consequently, the presented work revealed a strong bias which is caused by the incomplete way of recording variable objects. Namely, if an object has been observed but not detected in this observation it is currently not noted in the database. This is a very poor way of recording observations as a non-detection, and therefore a lower limit in terms of magnitude, can yield extremely valuable information about the nature of the object. On the other hand, the missing recording of these events gives rise to an misleading increase in performance accuracy by including heavily biased features such as the *StetsonK* value. While the *StetsonK* value is subject to observational biases, other values like the *PercentAmplitude* and *PercentDifferenceFluxPercentile (PDFP)* include effects of spatial bias. The distance modulus ($dm = m - M$) towards the Magellanic clouds is of the order of $dm_{\mathrm{MC}} = 18.5$. Consequently, the amplitude of

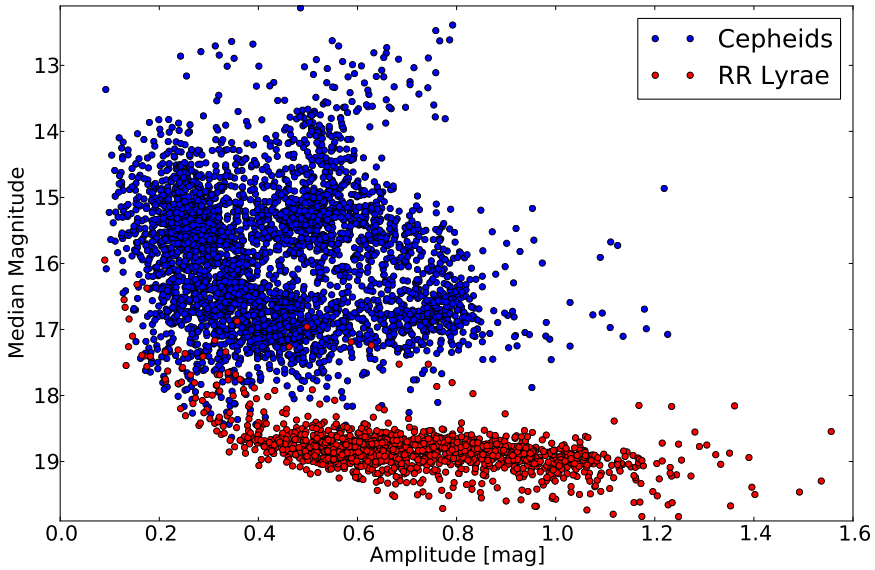
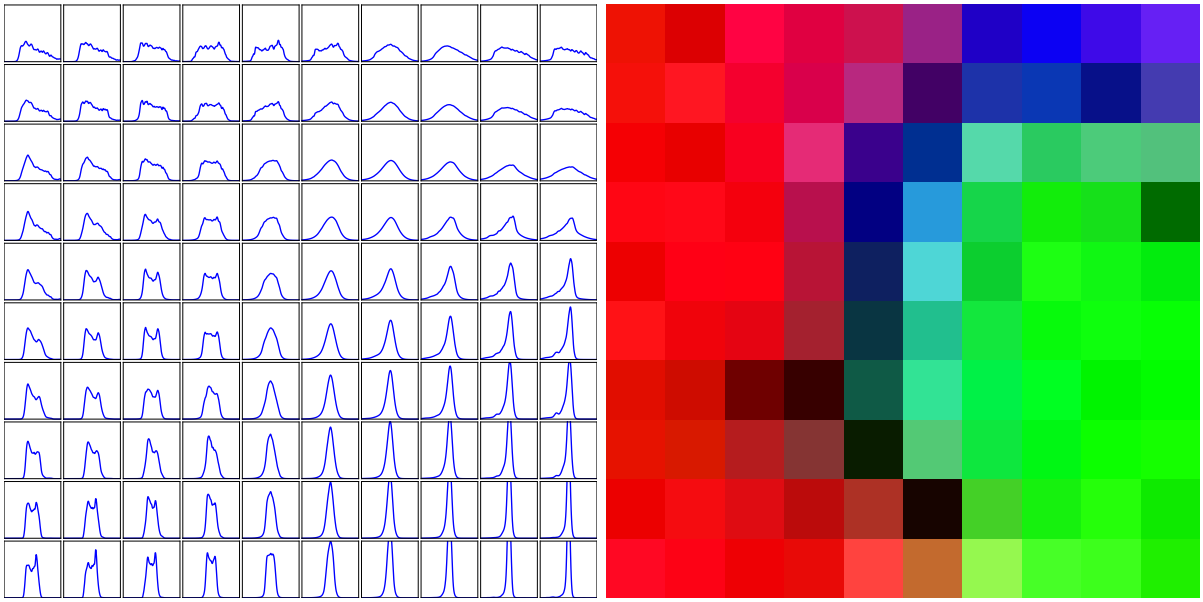


Figure 4.4: *The clear separation between cepheids and RR Lyrae stars in the OGLE dataset only exists, since all stars are located in the Magellanic clouds. However, this is a very special situation which does not generalize for all regions of the sky, therefore the use of features including the median brightness is highly questionable.*

variability and the apparent magnitudes provide quite strong evidence on the location of the observed object. As an example, cepheids and RR Lyrae stars both vary typically with an amplitude up to 1.5 mag, therefore they would be indistinguishable using just the amplitude. On the other hand, OGLE is focusing on the observation of the Magellanic clouds and thus the apparent brightness of RR Lyrae (cepheids) stars is of the order of 19 mag (16 mag). In Figure 4.4, it is shown that the introduction of the apparent magnitude provides additional information allowing for a better distinction between the classes. However, it is questionable whether such incidental knowledge should be used in the classification of light curves, especially when the survey observes large areas in the sky which do not obey this behavior. If the same classification scheme would now be used for an all-sky survey (such as ASAS) where the majority of stars is observed in the Milky Way, the learned classifier would not be applicable. If the median apparent magnitude is used, it should be used as a singular feature and not be encoded with other characteristics as done in *PercentAmplitude* and *PDFP*.

Another great advantage of the newly introduced methodology is that now even unsupervised algorithms may be applied in a meaningful way. This is of course also possible for a given set

of features, however, removing or changing the importance of individual features will change the outcome of the algorithm significantly. The presented representation is a very natural one and it has been shown that the $L2$-norm is a suitable measure to describe distances between the probability densities. Therefore, a self-organizing map (SOM, also called Kohonen map, Kohonen, 1990) is suited to visualize the prototypes which are inherent in a given dataset. In Figure 4.5a the Kohonen-map for the static probability densities of the OGLE data can be seen. Each of the data items provided to the SOM is normalized, the learned prototypes are, however,



(a) *Trained self-organizing map* **(b)** *Class memberships*

Figure 4.5: *A self-organizing map trained on the static probability densities of the OGLE data. The obtained prototypes are shown in (a), in (b) the class assignment for Cepheids (red), binaries (green) and RR Lyrae stars (blue) are shown.*

not necessarily normalized. One can see in Figure 4.5b that all of the obtained prototypes can be more or less uniquely associated with one of the provided classes. Note that the frequency with which a class occurs has a considerable impact on the area covered in the self-organizing map. Therefore, the RR Lyrae stars (blue) are confined in a small region. This effect could be avoided by changing the distance and learning function of the SOM, this would, however, not change the general conclusions drawn from it in the following. For datasets with more classes (such as ASAS), more confusion is also added to the SOM. The classes will then be not as separated anymore since only the static behavior is considered and different classes become distinct only if the time behavior is included. However, the presented methodology can provide a rough guess on the class of a given light curve. Subsequent inclusion of the time behavior, can then describe the different classes in more detail.

Future work The most obvious extension to the density-based classification algorithm is to include the temporal behavior. Again, the explicitly time-dependent features can be strongly survey-dependent. Therefore, the included features have to be shown to be independent of sampling and observation strategy! The only defining temporal property which is potentially unbiased would be the estimate of the period. It should be noted that all the period-finding algorithms do depend on the sampling and require a large number of independent observations. Therefore, the quality of the period estimate is depending on the sampling and on the photo-

metric uncertainties of the observations. Furthermore, the chosen period-finding algorithm also has an impact. Therefore, even the period and its uncertainty (or false alarm probability) can be only included with care. The inclusion of the period is not entirely straight-forward, and two obvious ways might be investigated. With the known periods of all objects, the distance matrix can be extended with respect to the period

$$D\left(i,j\right) = D^{static}\left(i,j\right) + \gamma D^{period}\left(i,j\right) \tag{4.1}$$

where D^{static} is the distance with respect to the static distribution and γ is a freely tunable parameter which has to be optimized. $D^{period}\left(i,j\right)$ is the distance in terms of period which could be computed by

$$D^{period}\left(i,j\right) = \frac{P_i}{P_j} \quad \textbf{or} \quad D^{period}\left(i,j\right) = log10\left(\frac{P_i}{P_j}\right) \tag{4.2}$$

with P_i being the period of data item i. Parameter γ does not need to be a scalar, but could be also a matrix which is adapted for the comparison between all classes. This optimization of the metric is known as *metric learning* (see Xing *et al.*, 2003, and references herein). The down-side of this approach is that yet another parameter has to be introduced. In order to circumvent this problem, the classification could be also run on the static matrix and the period-matrix D^{period} independently. In order to obtain the final class label, the respective classifier should be able to return a probability so that eventually a *Naive Bayes classifier* (Rish, 2001) can be applied. Apparently, this approach cannot be fine-tuned with an extra parameter and comes at the cost of lower flexibility. In addition to the inclusion of the period, the generality of the concept can be tested by performing transfer learning with the proposed algorithm. In this approach the classifier would be trained on one dataset (e.g., ASAS) and is supposed to provide labels on another one (e.g., OGLE). If the applicability of this transfer learning scheme was validated, the classifier can be also trained on a mixed dataset. This would provide a good alternative to visual inspection for new upcoming surveys. The knowledge of a visually classified subsample can then be included as well and thereby enlarge the knowledge which is made available to the classifier.

Chapter 5

Summary

In this thesis, state-of-the-art machine learning techniques were applied to complex and high-dimensional databases. Machine learning approaches imply that no explicit physical model was employed on the data, instead general models (e.g., ESN) or the data themselves have been used as a model (redshift regression, light curve classification). In order to apply these learning techniques, a new fixed-length vector representation for the data was obtained. Subsequently, the distance/similarity between the new representations of the provided data was measured. This similarity measure is then used to obtain regression values, visualize , and classify test observations. The big advantage of analyzing datasets without using explicit models is that the analysis is not influenced by potentially wrong (or just incomplete) models. Additionally, data-driven algorithms are more flexible and able to learn from the provided dataset. The model is thus evolving with the supplied data and the decision about the emergence of a new class is effectively left to an objective function. In this way, the approach is less biased more easily reproducible than the human decision making. Therefore, learning approaches provide a powerful alternative to the manual inspection which is still the most prominent approach in astronomy. The power, inherent in data-driven approaches, was shown in three very distinct science cases which cover a broad range of common astronomical tasks, namely regression, visualization, and classification. In addition, the respective methods were applied on very different data types, covering optical spectra as well as regularly and irregularly sampled light curves0. The major focus of this work was to provide new representations of astronomical data that are flexible and can be easily extended to a broad range of similar observations and problems. Subsequently, these new representations were made available to state-of-the-art learning algorithms to achieve the required tasks. In order to enable the astronomical community to use the presented methodology in diverse science cases, it is necessary that the methodology is easily applicable, i.e., has a low number of freely tunable parameters, and is computationally efficient.

Redshift estimation The detection of SMBHB, gravitational lenses and other peculiar objects is an actively debated topic in astronomy. For this purpose, SDSS spectra were inspected for the existence of potential multiple redshift systems. Existing approaches focused thereby on shifts between pre-defined spectral features, e.g., double-peak $[OIII]$-lines (Smith *et al.*, 2010) or shifts between the $[OIII]$ and the H_β (Tsalmantza *et al.*, 2011). However, focusing only on predefined spectral features limits the generality, and thereby the completeness, of these approaches. The results from the SDSS redshift pipeline are in this case of little use since the analysis is not tailored for the detection of objects not describable by any of the template prototypes.

In the presented work, an approach that can detect multiple redshift systems for different spectral features and independent of the underlying continuum was introduced. This was done

by first subtracting the continuum to obtain a new vector representation tailored for estimating redshifts. The redshift of a test spectrum was estimated by scanning the database for the most similar representations using a k nearest neighbor search. Eventually, the median of their redshifts was then assigned to the test spectrum. For matters of clarity, the analysis was solely performed on a subset of $\approx 16,000$ SDSS spectra. Within those $\approx 0.1\%$ showed peculiarities in their spectra which were either caused by badly subtracted night sky or by the existence of multiple redshifts. A big advantage of the presented methodology over the SDSS redshift pipeline is that the simultaneous fitting of the class and the redshift can be disentangled. This led to the discovery of SDSS J120419.07-001855.93, an E+A galaxy with a spectrum that is very similar to that of a single F-star. The presented methodology does not respect the continuum behavior of the spectrum and does not get confused by the underlying blackbody emission of the spectrum. On the contrary, the SDSS pipeline cannot separate class and redshift fitting and therefore fails to describe the spectrum since only non-stellar sources are tested for larger redshifts. Hence, the methodology can be used as an alternative cross-validation for the results obtained by SDSS. With ever growing spectral databases, an independent estimate of the redshift, without imposing any explicit physical model, will gain importance as it allows to cross-validate the correctness and detect outliers at the same time.

Time series visualization Time series, obtained from space-based observations, provide an unhampered and complete view on the dynamical behavior of astronomical sources. Surveys, such as Kepler, provide an immense amount of data and therefore enable astronomers to reveal formerly unknown variability behavior. So far, the detection and analysis of the Kepler light curves focused on the search for dedicated object classes with already known variability behavior (e.g., RR Lyrae in Kolenberg *et al.*, 2010 or binaries in Prša *et al.*, 2011). In a more general classification by Debosscher *et al.* (2011), more than 60% of the sources are not assigned to any of the pre-defined classes. Apart from this, the reliability of the assigned labels is highly questionable. For example, if the catalog by Prša *et al.*, 2011 is used as ground truth, the precision of Debosscher *et al.* (2011) in detecting binaries is $< 50\%$. Thus, a more general approach to visualize (and later on classify) all light curves according to their latent dynamics bears a huge discovery potential.

In the second publication, visualization was performed by coupling an echo-state network with an autoencoder. The ESN captures the latent dynamics inherent in the light curves by respecting sequential properties, such as invariance against shifts and the variable length of sequences. The autoencoder is then used to reduce the dimensionality of the parameters resulting from the ESN to two. The crux of this algorithm is that, due to the coupling, the reconstruction error caused by the encoding can be directly measured on the data, instead of merely measuring the reconstruction error on the model parameters. For a set of labeled artificial and X-ray data, it has been shown that the coupling between the encoder and the model allows to separate different variability classes in a more reliable way. The subsequent application of the newly developed methodology to 300 unclassified Kepler stars allowed to cluster the variable sources according to their dynamical behavior. Even though, the visualization quality of the presented algorithm on unclassified Kepler data was judged only empirically, it is satisfying to see that different subclasses with different dynamical properties emerge. The obvious next step would be to apply the algorithm to the entire Kepler database. A step towards this direction was made in Kügler *et al.* (2015) where $\approx 6,200$ non-periodic Kepler light curves were visualized. Only being presented the light curves, the visualization appears to learn about physical properties (e.g., surface gravity) and separates eventually main sequence stars from giants in the visualization. While this behavior has been noticed before (Huber *et al.*, 2010; Bastien *et al.*, 2013), it is pleasing to see that this correlation is also revealed by the presented visualization algorithm.

Light curve classification Ground-based observations of variable sources are sampled very irregularly. The incompleteness of the observation is depending on seasonal effects, the weather at a given location, and the observation strategy of a given survey. Consequently, the obtained light curves are extremely inhomogeneous. As stated in the former paragraph, comparing (even regularly sampled) light curves is not straight-forward; the inhomogeneity adds another uncertainty because now even the knowledge about the dynamical behavior is incomplete. The common strategy to represent a given light curve as a fixed-length vector is the extraction of features. This representation can be used to train classifiers on a subset of light curves that have been manually labeled. In Richards *et al.* (2011) it has been argued that as many features as possible should be extracted and the decision about the importance of features can be left to the very powerful random forest classifier. This approach has shown to be reliable on different datasets, however, other issues relating to the use of features have been neglected so far. First of all, the features should not depend directly on properties that are survey-inherent, e.g., sampling. It can be easily shown that some of the features[17] do not obey this behavior and are strongly biased due to the observation (recording) strategy. Secondly, the features are treated as point estimates which strongly limits their use if they are extracted from different surveys. Even within a single survey, the uncertainty can vary significantly as fainter sources tend to show higher uncertainties. Lastly, the feature-based classification does not allow to infer any physical knowledge about mis-classified sources as (most of) the features are meaningless from an astrophysical perspective. The arbitrariness of the feature selection, and therefore of the distances between light curves, prohibits the use of unsupervised methodology as the outcome would differ drastically if features are added or omitted.

The introduced methodology uses probability density functions (PDFs) to represent the static part of the light curves. It was shown that the classification performance is comparable to that of state-of-the-art classifiers (using the random forest) if only *static* features are considered. Besides that, the use of PDFs has great advantages. Since the data representation is a very natural one, the only degrees of freedom in the presented methodology are the choice of the metric and of the classifier. In contrast to that, the feature-based classification is preceded by feature selection and normalization. Another disadvantage of the features is, that they cannot deal properly with outliers, while in the density-based approach only a low probability is added to the overall distribution which, depending on the chosen metric, is negligible. Additionally, the photometric uncertainties are captured accordingly and therefore also non-detections can be included quite naturally, if recorded. Finally, the new density-based representation allows a meaningful application of unsupervised methodology, such as the self-organizing map shown in Section 4.3.

Conclusion The findings of this thesis highlight that new approaches are needed to solve the mentioned scientific questions. The advantages of the introduced methodology were highlighted in very diverse, astronomically relevant science cases. The application of the approaches to huge databases revealed interesting objects on one side and severe drawbacks of the existing methodology on the other. However, this implies by *no means* that the data-driven approaches are always superior and should replace existing ones. The presented algorithms should be seen as complementary approaches which are of a high value when big volumes of data are available. In that case, data-driven algorithms can provide a significantly less biased view on the data which is especially helpful when the physics behind these high-dimensional and highly complex data still needs to be explored. Additionally, the application of a well designed machine learning approach to a large dataset should recover the knowledge independently obtained by human inspection.

[17] In the presented case, it has been only shown for static features. This statement holds also for non-static features, e.g, the maximum slope.

However, it is important to notice that it is much harder to understand the results of a machine learning task than the ones of a model-driven approach. A model-driven approach allows a direct interpretation of physical properties rather than incomprehensible (and also physically meaningless) parameters. Fixing the missing link between meaningless parameters which originate from the application of purely data-driven methods and real physical quantities should be the main concern on the data analyst's and astronomer's side as only this can eventually lead to a common understanding of data. A step towards this goal has been done by applying physical models to existing data in a probabilistic setting. For example, Lewis and Bridle (2002) provide a fully probabilistic estimate of the parameters of the cosmic microwave background (CMB). In parallel, the new direction of semi-supervised machine learning approaches is emerging slowly (see, e.g., Richards *et al.*, 2012). There, substantial but incomplete knowledge can be provided to the algorithm and eventually the best solution, balancing the given knowledge and evidence acquired from the data, is found.

It was shown, that the way of representing (and also preprocessing) the available astronomical data has a huge impact on the usability of data-driven approaches. Effectively, the new representation is translating the scientific question into a form that is understandable by learning approaches. This is quite nicely illustrated in the first publication. If only the raw data would be fed to a regression algorithm, the similarity of spectra would be dominated by the continuum and therefore the similarity would not be judged according to redshift. The representation of the data is thus a trade-off between making the data understandable for the learning approach and keeping them meaningful for scientists. For visualizing light curves, a purely data-driven model was employed to obtain a new representation. While the resulting weights of the ESN are entirely useless for a physical interpretation, the coupling of the autoencoder to the ESN allows to study the impact of a change in the weights on modeling the light curves. This enables us to obtain a physical intuition about the similarity of light curves in a two-dimensional projection, even though one has to be still aware of the fact that the underlying ESN interprets the similarity between light curves potentially quite different than our human perception would. Therefore, the coupling of a physically motivated model to the autoencoder is a valuable method to obtain an understanding of the latent behavior. Eventually, by classifying light curves it has been shown that understanding the extracted features is way more troublesome than understanding the PDF. On the other hand, learning approaches are tailored for vectorial data (features) and therefore a broader variety of algorithms exists, for example the random forest, that are not applicable for this more general representation.

Conclusively, the application of machine learning approaches in astronomy is still in its infancy. The potential, in terms of science cases, to be tackled with alternate methodology is huge. The analysis of radio cubes (e.g., Punzo *et al.*, 2015), morphological studies (e.g., Lintott *et al.*, 2008), the automated clustering of spectral databases (e.g., Cui *et al.*, 2012) and the treatment of upcoming (peta-byte sized) sequential databases (e.g., Ivezi *et al.*, 2011) are certainly just the most prominent examples and reflect only the tip of the iceberg. It is inevitable that computer scientists, statisticians and astronomers work hand in hand so that astronomy can benefit from the knowledge provided by these disciplines.

Bibliography

Ahn, C. P., Alexandroff, R., Allende Prieto, C., *et al.* (2014). The Tenth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the SDSS-III Apache Point Observatory Galactic Evolution Experiment. *The Astrophysical Journal Supplement Series*, **211**, 17.

Antonucci, R. (1993). Unified models for active galactic nuclei and quasars. *Annual Review of Astronomy and Astrophysics*, **31**, 473–521.

Auvergne, M., Bodin, P., Boisnard, L., *et al.* (2009). The CoRoT satellite in flight: description and performance. *Astronomy and Astrophysics*, **506**, 411–424.

Ball, N. M. and Robert, R. J. (2010). DATA MINING AND MACHINE LEARNING IN ASTRONOMY. *International Journal of Modern Physics D*, **19**, 1049–1106.

Bastien, F. A., Stassun, K. G., Basri, G., and Pepper, J. (2013). An observational correlation between stellar brightness variations and surface gravity. *Nature*, **500**, 427–430.

Becker, R. H., White, R. L., and Helfand, D. J. (1995). The FIRST Survey: Faint Images of the Radio Sky at Twenty Centimeters. *The Astrophysical Journal*, **450**, 559.

Begelman, M. C., Blandford, R. D., and Rees, M. J. (1980). Massive black hole binaries in active galactic nuclei. *Nature*, **287**, 307–309.

Belloni, T., Klein-Wolt, M., Méndez, M., van der Klis, M., and van Paradijs, J. (2000). A model-independent analysis of the variability of GRS 1915+105. *Astronomy and Astrophysics*, **355**, 271–290.

Benedict, G. F., McArthur, B. E., Feast, M. W., *et al.* (2007). Hubble Space Telescope Fine Guidance Sensor Parallaxes of Galactic Cepheid Variable Stars: Period-Luminosity Relations. *Astronomical Journal*, **133**, 1810–1827.

Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Commun. ACM*, **18**(9), 509–517.

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

Bolton, A. S., Burles, S., Koopmans, L. V. E., Treu, T., and Moustakas, L. A. (2006). The Sloan Lens ACS Survey. I. A Large Spectroscopically Selected Sample of Massive Early-Type Lens Galaxies. *The Astrophysical Journal*, **638**, 703–724.

Bonnarel, F., Fernique, P., Bienaymé, O., *et al.* (2000). The ALADIN interactive sky atlas. A reference tool for identification of astronomical sources. *Astronomy and Astrophysics Supplement*, **143**, 33–40.

Borucki, W. J., Koch, D., Basri, G., *et al.* (2010). Kepler Planet-Detection Mission: Introduction and First Results. *Science*, **327**, 977–.

Breiman, L. (2001). Random forests. *Machine Learning*, pages 5–32.

Cáceres, C. and Catelan, M. (2008). The Period-Luminosity Relation of RR Lyrae Stars in the SDSS Photometric System. *The Astrophysical Journal Supplement Series*, **179**, 242–248.

Chang, C.-C.and Lin, C.-J. (2011). LIBSVM : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*.

Collinge, M. J., Strauss, M. A., Hall, P. B., *et al.* (2005). Optically Identified BL Lacertae Objects from the Sloan Digital Sky Survey. *Astronomical Journal*, **129**, 2542–2561.

Cook, K. H., Alcock, C., Allsman, H. A., *et al.* (1995). Variable Stars in the MACHO Collaboration Database. In R. S. Stobie and P. A. Whitelock, editors, *IAU Colloq. 155: Astrophysical Applications of Stellar Pulsation*, volume 83 of *Astronomical Society of the Pacific Conference Series*, page 221.

Cui, X.-Q., Zhao, Y.-H., Chu, Y.-Q., *et al.* (2012). The Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST). *Research in Astronomy and Astrophysics*, **12**, 1197–1242.

Debosscher, J., Sarro, L. M., Aerts, C., *et al.* (2007). Automated supervised classification of variable stars. I. Methodology. *Astronomy and Astrophysics*, **475**, 1159–1183.

Debosscher, J., Blomme, J., Aerts, C., and De Ridder, J. (2011). Global stellar variability study in the field-of-view of the Kepler satellite. *Astronomy and Astrophysics*, **529**, A89.

Decarli, R., Dotti, M., Montuori, C., Liimets, T., and Ederoclite, A. (2010). The Peculiar Optical Spectrum of 4C+22.25: Imprint of a Massive Black Hole Binary? *The Astrophyiscal Journal Letters*, **720**, L93–L96.

Donalek, C., Kumar, A. A., Djorgovski, S. G., *et al.* (2013). Feature Selection Strategies for Classifying High Dimensional Astronomical Data Sets. *preprint, (arXiv:1310.1976)*.

Drake, A. J., Djorgovski, S. G., Mahabal, A., *et al.* (2009). First Results from the Catalina Real-Time Transient Survey. *The Astrophysical Journal*, **696**, 870–884.

Dressler, A. and Gunn, J. E. (1983). Spectroscopy of galaxies in distant clusters. II - The population of the 3C 295 cluster. *The Astrophysical Journal*, **270**, 7–19.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press*, page 226â231.

Fu, H., Yan, L., Myers, A. D., *et al.* (2012). The Nature of Double-peaked [O III] Active Galactic Nuclei. *The Astrophysical Journal*, **745**, 67.

Gautschy, A. and Saio, H. (1995). Stellar Pulsations Across The HR Diagram: Part 1. *Annual Review of Astronomy and Astrophysics*, **33**, 75–114.

Gillessen, S., Eisenhauer, F., Trippe, S., *et al.* (2009). Monitoring Stellar Orbits Around the Massive Black Hole in the Galactic Center. *The Astrophysical Journal*, **692**, 1075–1109.

Gilliland, R. L., Brown, T. M., Christensen-Dalsgaard, J., *et al.* (2010). Kepler Asteroseismology Program: Introduction and First Results. *Publications of the Astronomical Society of the Pacific*, **122**, 131–143.

Goto, T. (2005). 266 E+A galaxies selected from the Sloan Digital Sky Survey Data Release 2: the origin of E+A galaxies. *Monthly Notices of the Royal Astronomical Society*, **357**, 937–944.

Graham, M. J., Drake, A. J., Djorgovski, S. G., *et al.* (2013). A comparison of period finding algorithms. *Monthly Notices of the Royal Astronomical Society*, **434**, 3423–3444.

Graham, M. J., Djorgovski, S. G., Drake, A. J., *et al.* (2014). A novel variability-based method for quasar selection: evidence for a rest-frame ∼54 d characteristic time-scale. *Monthly Notices of the Royal Astronomical Society*, **439**, 703–718.

Graham, M. J., Djorgovski, S. G., Stern, D., *et al.* (2015). A possible close supermassive black-hole binary in a quasar with optical periodicity. *Nature*, **518**, 74–76.

Greiner, J., Morgan, E. H., and Remillard, R. A. (1996). Rossi X-Ray Timing Explorer Observations of GRS 1915+105. *The Astrophysical Journal Letters*, **473**, L107.

Gualandris, A. and Merritt, D. (2008). Ejection of Supermassive Black Holes from Galaxy Cores. *The Astrophysical Journal*, **678**, 780–797.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). Data Mining, Inference, and Prediction. In *The Elements of Statistical Learning Second Edition*. Springer.

Heidt, J. and Wagner, S. J. (1996). Statistics of optical intraday variability in a complete sample of radio-selected BL Lacertae objects. *Astronomy and Astrophysics*, **305**, 42.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**, 417–441, 498–520.

Huber, D., Bedding, T. R., Stello, D., *et al.* (2010). Asteroseismology of Red Giants from the First Four Months of Kepler Data: Global Oscillation Parameters for 800 Stars. *ApJ*, **723**, 1607–1617.

Ivezi, Z., Tyson, J., Acosta, E., *et al.* (2011). LSST: FROM SCIENCE DRIVERS TO REFERENCE DESIGN AND ANTICIPATED DATA PRODUCTS. *arXiv*, **0805**, 2366.

Ivezić, Ž., Smith, J. A., Miknaitis, G., *et al.* (2007). Sloan Digital Sky Survey Standard Star Catalog for Stripe 82: The Dawn of Industrial 1% Optical Photometry. *Astronomical Journal*, **134**, 973–998.

Ivezić, Ž., Connolly, A., VanderPlas, J., and Gray, A. (2014). *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data*. Princeton Series in Modern Observational Astronomy.

Kannan, R., Macciò, A. V., Fontanot, F., *et al.* (2015). From discs to bulges: effect of mergers on the morphology of galaxies. *Monthly Notices of the Royal Astronomical Society*, **452**, 4347–4360.

Kochanek, C. S. (1991). The implications of lenses for galaxy structure. *The Astrophysical Journal*, **373**, 354–368.

Kochanek, C. S. (2002). What Do Gravitational Lens Time Delays Measure? *The Astrophysical Journal*, **578**, 25–32.

Kohonen, T. (1990). The self-organizing map. *IEEE*, **78 (9)**, 1464–1480.

Kolenberg, K., Szabó, R., Kurtz, D. W., *et al.* (2010). First Kepler Results on RR Lyrae Stars. *The Astrophysical Journal Letters*, **713**, L198–L203.

Komossa, S., Burwitz, V., Hasinger, G., *et al.* (2003). Discovery of a Binary Active Galactic Nucleus in the Ultraluminous Infrared Galaxy NGC 6240 Using Chandra. *The Astrophysical Journal*, **582**, L15–L19.

Kramer, M. A. (1991). Non-linear principal component analysis using autoassociative neural networks. *AICHE Journal*, **37**, 233–243.

Kügler, S. D., Gianniotis, N., and Polsterer, K. L. (2015). An Explorative Approach for Inspecting Kepler Data. *ArXiv e-prints*.

Lewis, A. and Bridle, S. (2002). Cosmological parameters from CMB and other data: A Monte Carlo approach. *Physical Review D*, **66**(10), 103511.

Lindgren, H., Lundstrom, I., and Stenholm, B. (1975). Short-periodic light variations in Wolf-Rayet stars. *Astronomy and Astrophysics*, **44**, 219–222.

Lintott, C. J., Schawinski, K., Slosar, A., *et al.* (2008). Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, **389**, 1179–1189.

Liu, X., Shen, Y., Bian, F., Loeb, A., and Tremaine, S. (2014). Constraining Sub-parsec Binary Supermassive Black Holes in Quasars with Multi-epoch Spectroscopy. II. The Population with Kinematically Offset Broad Balmer Emission Lines. *The Astrophysical Journal*, **789**, 140.

Lloyd, S. (1982). "least squares quantization in pcm". *IEEE Transactions on Information Theory*, **28**(2), 129–137.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif. University of California Press.

Maness, H. L., Taylor, G. B., Zavala, R. T., Peck, A. B., and Pollack, L. K. (2004). Breaking All the Rules: The Compact Symmetric Object 0402+379. *The Astrophysical Journal*, **602**, 123–134.

Marshall, P. J., Treu, T., Melbourne, J., *et al.* (2007). Superresolving Distant Galaxies with Gravitational Telescopes: Keck Laser Guide Star Adaptive Optics and Hubble Space Telescope Imaging of the Lens System SDSS J0737+3216. *The Astrophysical Journal*, **671**, 1196–1211.

Marshall, S., Akerlof, C., Kehoe, R., *et al.* (1997). The ROTSE Project. In *American Astronomical Society Meeting Abstracts*, volume 29 of *Bulletin of the American Astronomical Society*, page 1290.

Matijevič, G., Prša, A., Orosz, J. A., *et al.* (2012). Kepler Eclipsing Binary Stars. III. Classification of Kepler Eclipsing Binary Light Curves with Locally Linear Embedding. *The Astronomical Journal*, **143**, 123.

Meusinger, H., Schalldach, P., Scholz, R.-D., *et al.* (2012). Unusual quasars from the Sloan Digital Sky Survey selected by means of Kohonen self-organising maps. *Astronomy and Astrophysics*, **541**, A77.

Milosavljević, M. and Merritt, D. (2003). The Final Parsec Problem. In J. M. Centrella, editor, *The Astrophysics of Gravitational Wave Sources*, volume 686 of *American Institute of Physics Conference Series*, pages 201–210.

Muñoz, J. A., Falco, E. E., Kochanek, C. S., *et al.* (1998). The Castles Project. *Astrophysics and Space Science*, **263**, 51–54.

Perryman, M. A. C., Lindegren, L., Kovalevsky, J., *et al.* (1997). The HIPPARCOS Catalogue. *Astronomy and Astrophysics*, **323**, L49–L52.

Perryman, M. A. C., de Boer, K. S., Gilmore, G., *et al.* (2001). GAIA: Composition, formation and evolution of the Galaxy. *Astronomy and Astrophysics*, **369**, 339–363.

Pojmanski, G. (1997). The All Sky Automated Survey. *ActaA*, **47**, 467–481.

Polsterer, K. L., Zinn, P.-C., and Gieseke, F. (2013). Finding new high-redshift quasars by asking the neighbours. *Monthly Notices of the Royal Astronomical Society*, **428**, 226–235.

Popović, L. Č. (2012). Super-massive binary black holes and emission lines in active galactic nuclei. *New Astronomy Reviews*, **56**, 74–91.

Protopapas, P., Giammarco, J. M., Faccioli, L., *et al.* (2006). Finding outlier light curves in catalogues of periodic variable stars. *Monthly Notices of the Royal Astronomical Society*, **369**, 677–696.

Prša, A., Batalha, N., Slawson, R. W., *et al.* (2011). Kepler Eclipsing Binary Stars. I. Catalog and Principal Characterization of 1879 Eclipsing Binaries in the First Data Release. *The Astronomical Journal*, **141**, 83.

Punzo, D., van der Hulst, J. M., Roerdink, J. B. T. M., *et al.* (2015). The role of 3-D interactive visualization in blind surveys of H I in galaxies. *Astronomy and Computing*, **12**, 86–99.

Richards, J. W., Starr, D. L., Butler, N. R., *et al.* (2011). On Machine-learned Classification of Variable Stars with Sparse and Noisy Time-series Data. *The Astrophysical Journal*, **733**, 10.

Richards, J. W., Homrighausen, D., Freeman, P. E., Schafer, C. M., and Poznanski, D. (2012). Semi-supervised learning for photometric supernova classification. *Monthly Notices of the Royal Astronomical Society*, **419**, 1121–1135.

Richstone, D., Ajhar, E. A., Bender, R., *et al.* (1998). Supermassive black holes and the evolution of galaxies. *Nature*, **395**, A14.

Rish, I. (2001). An empirical study of the naive bayes classifier. In *AIJCAI workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46.

Rodriguez, C., Taylor, G. B., Zavala, R. T., Pihlström, Y. M., and Peck, A. B. (2009). H I Observations of the Supermassive Binary Black Hole System in 0402+379. *The Astrophysical Journal*, **697**, 37–44.

Russell, H. N. (1914). Relations Between the Spectra and Other Characteristics of the Stars. *Popular Astronomy*, **22**, 275–294.

Samus, N. N., Durlevich, O. V., and et al. (2004). VizieR Online Data Catalog: Combined General Catalogue of Variable Stars (Samus+ 2004). *VizieR Online Data Catalog*, **2250**, 0.

Sanders, D. B., Soifer, B. T., Elias, J. H., *et al.* (1988). Ultraluminous infrared galaxies and the origin of quasars. *The Astrophysical Journal*, **325**, 74–91.

Scargle, J. D. (1982). Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal*, **263**, 835–853.

Schwarzenberg-Czerny, A. (1989). On the advantage of using analysis of variance for period search. *Monthly Notices of the Royal Astronomical Society*, **241**, 153–165.

Shakura, N. I. and Sunyaev, R. A. (1973). Black holes in binary systems. Observational appearance. *AAp*, **24**, 337–355.

Skrutskie, M. F., Cutri, R. M., Stiening, R., *et al.* (2006). The Two Micron All Sky Survey (2MASS). *Astronomical Journal*, **131**, 1163–1183.

Smith, K. L., Shields, G. A., Bonning, E. W., McMullen, C. C., Rosario, D. J., and Salviander, S. (2010). A Search for Binary Active Galactic Nuclei: Double-peaked [O III] AGNs in the Sloan Digital Sky Survey. *The Astrophysical Journal*, **716**, 866–877.

Soszyñski, I., Udalski, A., Szymañski, M. K., *et al.* (2009). The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. IV. Long-Period Variables in the Large Magellanic Cloud. *ActaA*, **59**, 239–253.

Stellingwerf, R. F. (1978). Period determination using phase dispersion minimization. *The Astrophysical Journal*, **224**, 953–960.

Stoughton, C., Lupton, R. H., Bernardi, M., *et al.* (2002). Sloan Digital Sky Survey: Early Data Release. *Astronomical Journal*, **123**, 485–548.

SubbaRao, M., Frieman, J., Bernardi, M., *et al.* (2002). The Sloan Digital Sky Survey 1-Dimensional Spectroscopic Pipeline. In J.-L. Starck and F. D. Murtagh, editors, *Astronomical Data Analysis II*, volume 4847 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 452–460.

Taylor, M. B. (2005). TOPCAT & STIL: Starlink Table/VOTable Processing Software. In P. Shopbell, M. Britton, and R. Ebert, editors, *Astronomical Data Analysis Software and Systems XIV*, volume 347 of *Astronomical Society of the Pacific Conference Series*, page 29.

Tiňo, P. and Raychaudhury, S. (2012). Computational Intelligence in Astronomy â A Win-Win Situation. In *Lecture Notes in Computer Science, Springer-Verlag, LNCS 7505*, pages 57–71.

Toomre, A. and Toomre, J. (1972). Galactic Bridges and Tails. *The Astrophysical Journal*, **178**, 623–666.

Tsalmantza, P., Decarli, R., Dotti, M., and Hogg, D. W. (2011). A Systematic Search for Massive Black Hole Binaries in the Sloan Digital Sky Survey Spectroscopic Sample. *The Astrophysical Journal*, **738**, 20.

Udalski, A., Szymanski, M., Kaluzny, J., Kubiak, M., and Mateo, M. (1992). The Optical Gravitational Lensing Experiment. *ActaA*, **42**, 253–284.

Urry, C. M. and Padovani, P. (1995). Unified Schemes for Radio-Loud Active Galactic Nuclei. *Publications of the Astronomical Society of the Pacific*, **107**, 803.

Valtonen, M. J., Lehto, H. J., Nilsson, K., *et al.* (2008). A massive binary black-hole system in OJ287 and a test of general relativity. *Nature*, **452**, 851–853.

van der Maaten, L. J. P. and Hinton, G. E. (2008). Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research*, **9**, 2579–2605.

Vanden Berk, D. E., Richards, G. T., Bauer, A., *et al.* (2001). Composite Quasar Spectra from the Sloan Digital Sky Survey. *Astronomical Journal*, **122**, 549–564.

Watson, C. L. (2006). The International Variable Star Index (VSX). *Society for Astronomical Sciences Annual Symposium*, **25**, 47.

White, S. D. M. and Rees, M. J. (1978). Core condensation in heavy halos - A two-stage theory for galaxy formation and clustering. *Monthly Notices of the Royal Astronomical Society*, **183**, 341–358.

Xing, E. P., Ng, A. Y., Jordan, M. I., and Russell, S. (2003). Distance metric learning, with application to clustering with side-information. In *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 15*, pages 505–512. MIT Press.

Zliobaite, I., Bifet, A., Gaber, M., Gabrys, B., Gama, J., Minku, L., and Musial, K. (2012). Next challenges for adaptive learning systems. *SIGKDD Explor. Newsl.*, **14**(1), 48–55.

Appendix A

SDSS Spectral Features

Feature	$\lambda_{\mathrm{vac}}[\mathring{A}]$	Type
O VI	1033.82	Em
Lyα	1215.24	Em/Abs
N V	1240.81	Em
O I	1305.53	Em
C II	1335.31	Em
Si IV	1397.61	Em
Si IV / O IV	1399.8	Em
C IV	1549.48	Em
He II	1640.4	Em
O III	1665.85	Em
Al III	1857.4	Em
C III	1908.734	Em
C II	2326.0	Em
Ne IV	2439.5	Em
Mg II	2799.117	Em
Ne V	3346.79	Em
Ne VI	3426.85	Em
O II - Doublet	3727.092, 3729.875	Em
He I	3889.0	Em
K	3934.777	Abs
H	3969.588	Abs
S II	4072.3	Em
Hδ	4102.89	Em/Abs
G	4305.61	Abs
Hγ	4341.68	Em/Abs
O III	4364.436	Em
Hβ	4862.68	Em/Abs
O III - Triplett	4932.603, 4960.295, 5008.24	Em

Table A.1: *List of spectral features used by SDSS to estimate the redshift of a given spectrum. The list can be found under http://classic.sdss.org/dr6/algorithms/linestable.html*

Feature	$\lambda_{\text{vac}}[\mathring{A}]$	Type
Mg	5176.7	Abs
Na	5895.6	Abs
O I - Doublet	6302.046, 6365.536	Em
N I	6529.03	Em
N II	6549.86	Em
Hα	6564.61	Em/Abs
N II	6585.27	Em
S II - Doublet	6718.29, 6732.67	Em
CaII - Triplett	8500.36, 8544.44, 8664.0	Abs

Table A.1 continued.

Appendix B

Redshift Regression for Individual Regions

Here the subplots for all chosen spectral regions are shown individually. The meaning of them is the same as the one in Figure 6 of the first publication.

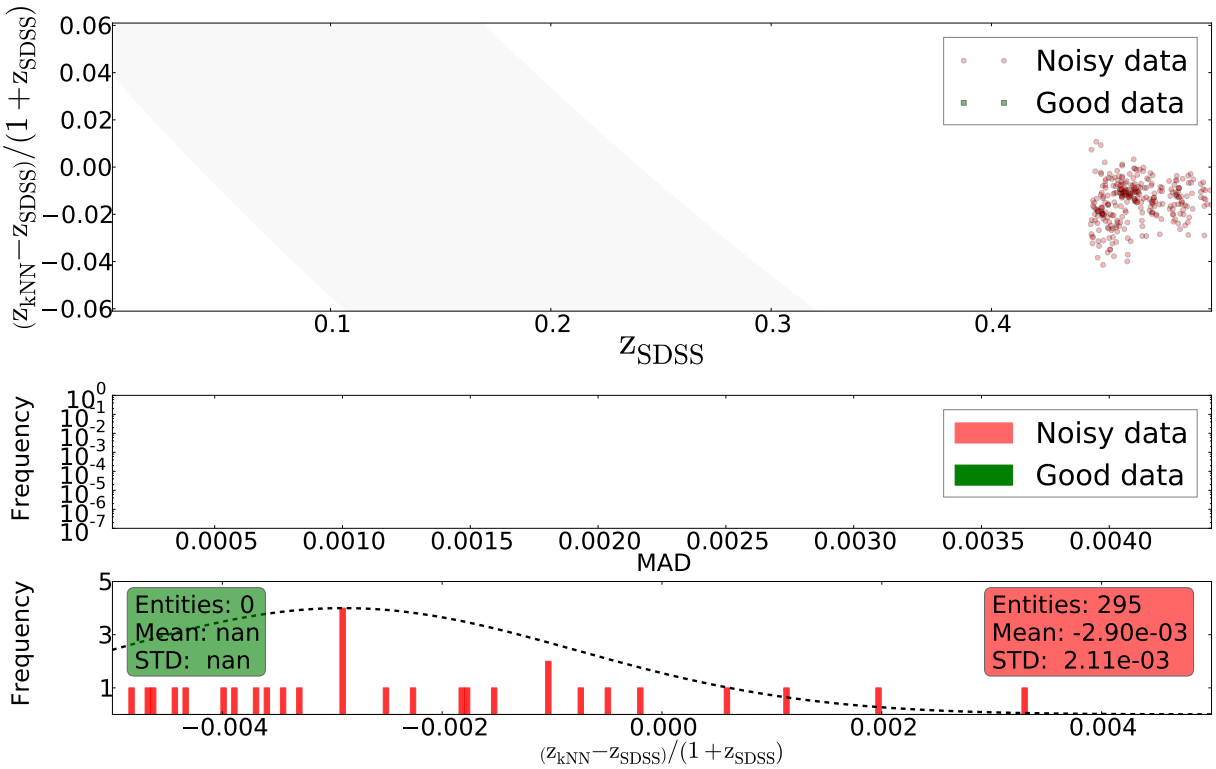


Figure B.1: *Regressional redshift versus spectroscopic one for the emission feature MgII*

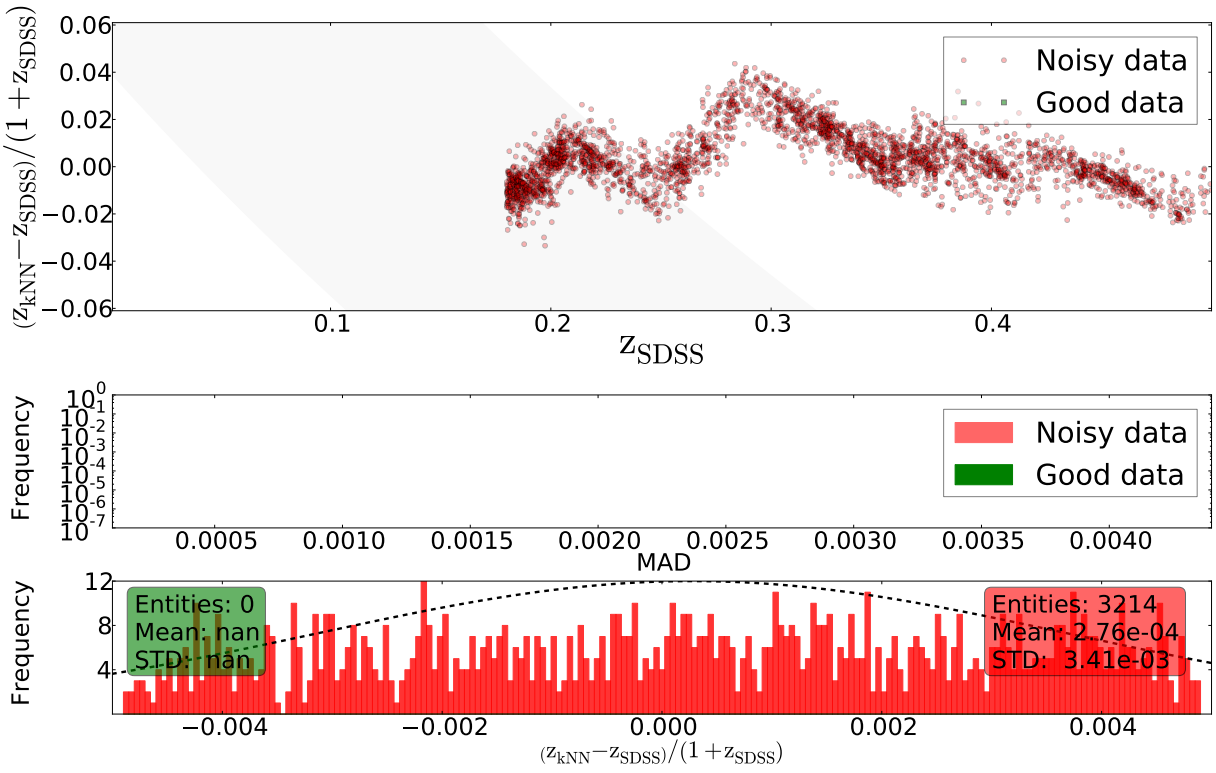


Figure B.2: *Regressional redshift versus spectroscopic one for the emission feature NeV*

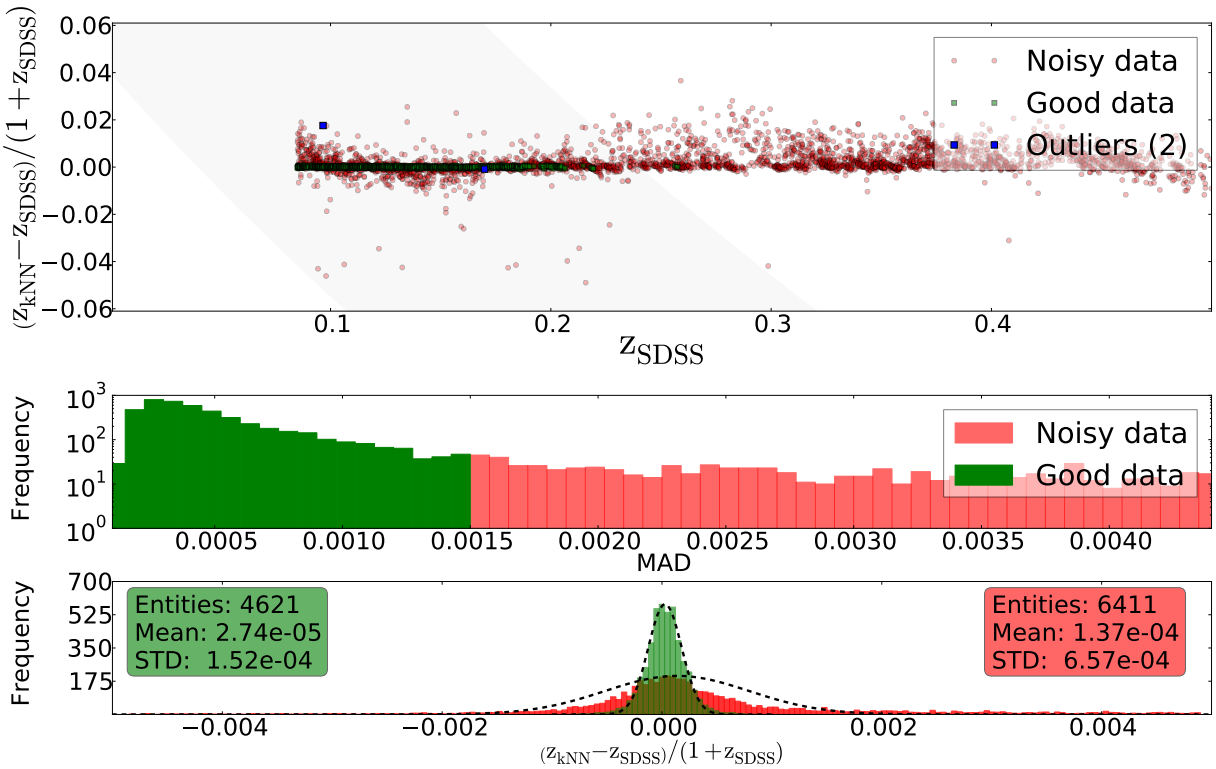


Figure B.3: *Regressional redshift versus spectroscopic one for the emission feature [OII]*

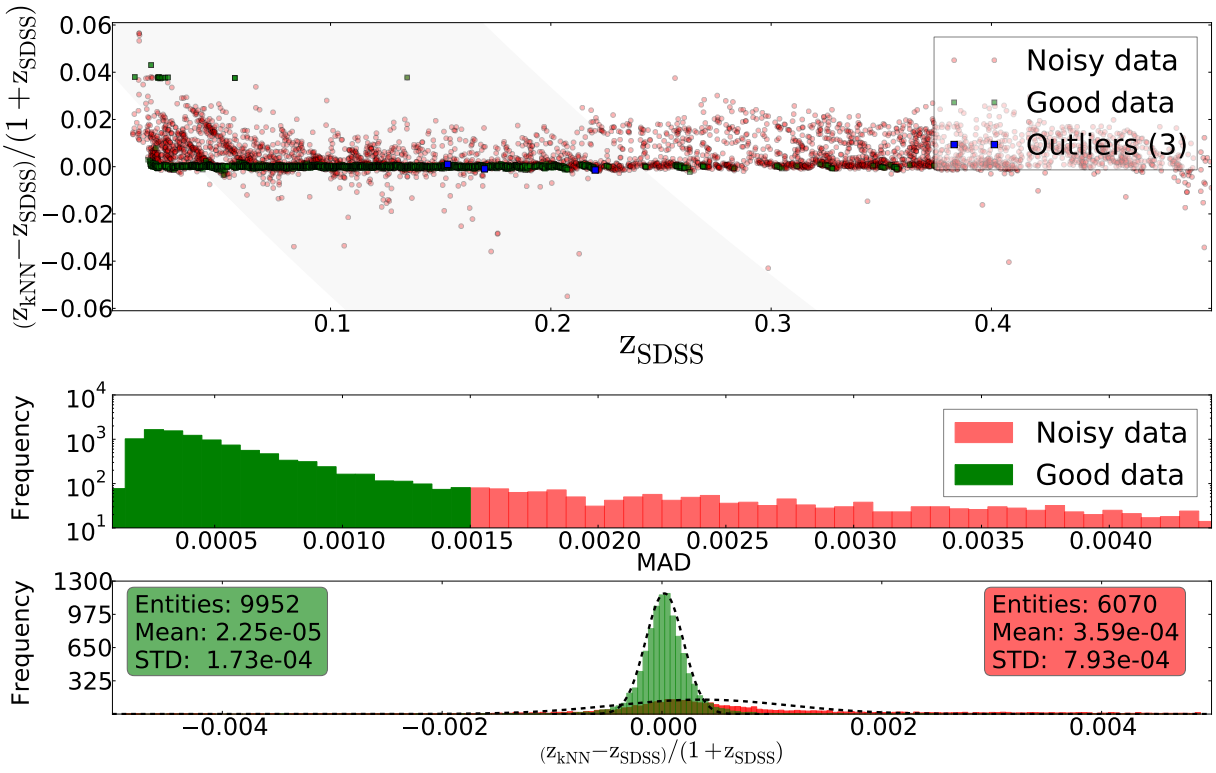


Figure B.4: *Regressional redshift versus spectroscopic one for the emission feature H_ϵ,H_ζ*

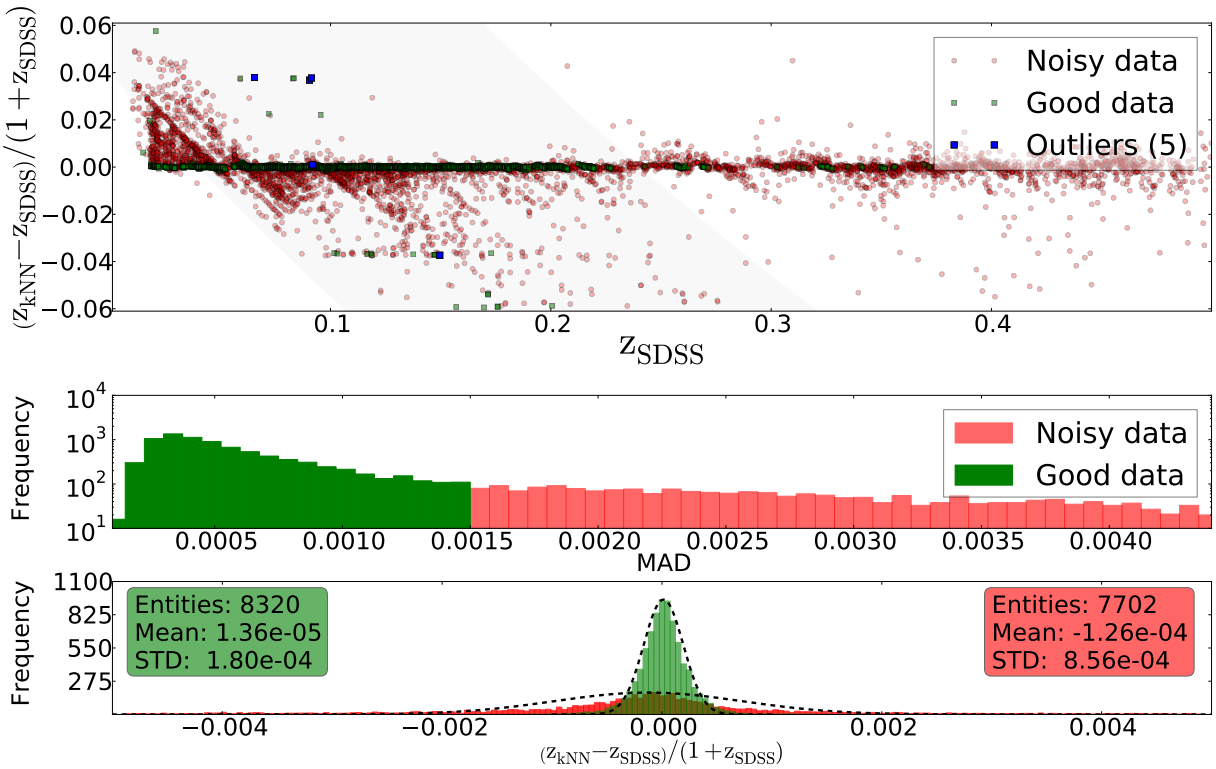


Figure B.5: *Regressional redshift versus spectroscopic one for the emission feature H_δ*

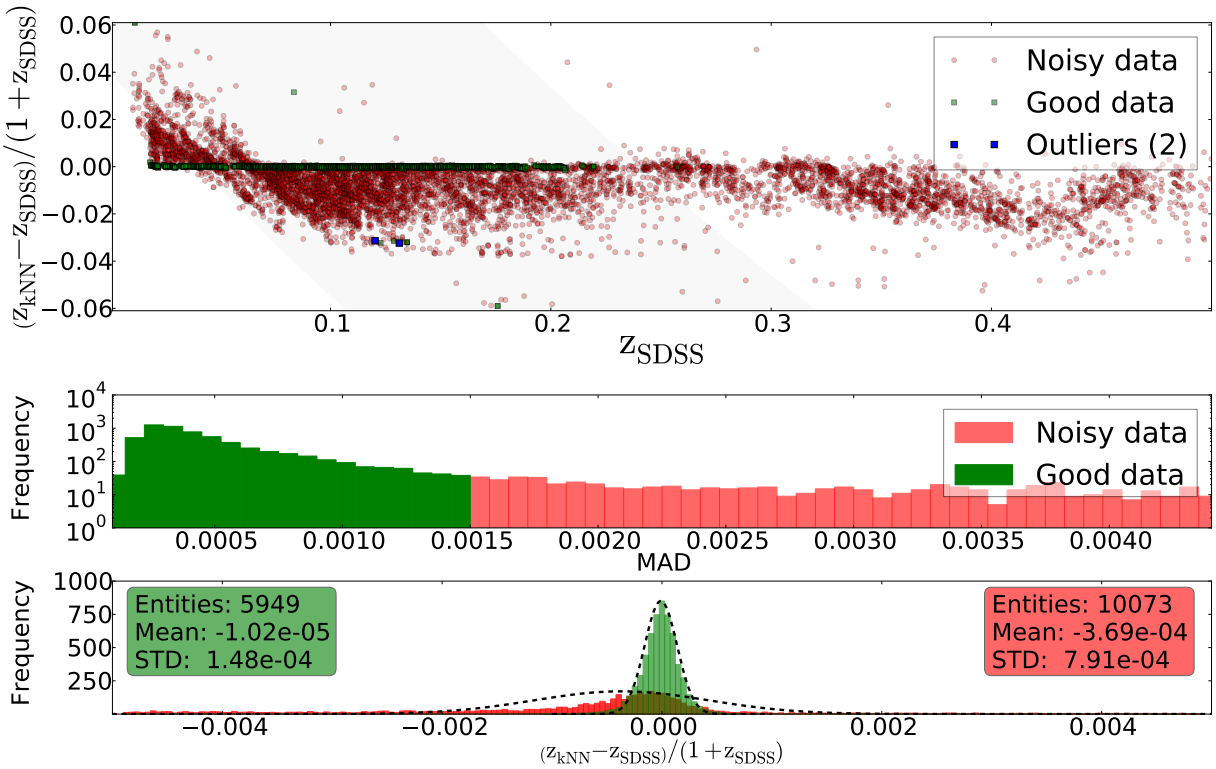


Figure B.6: *Regressional redshift versus spectroscopic one for the emission feature H_γ*

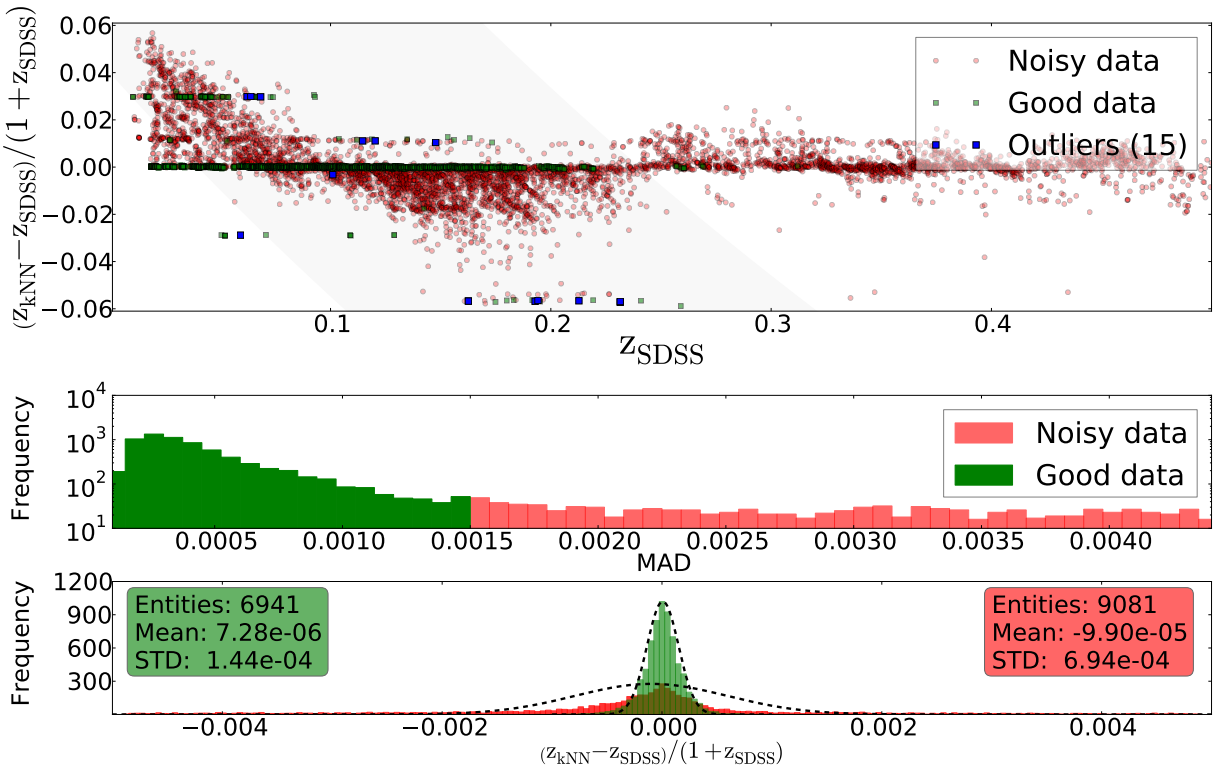


Figure B.7: *Regressional redshift versus spectroscopic one for the emission feature H_β, [OIII]*

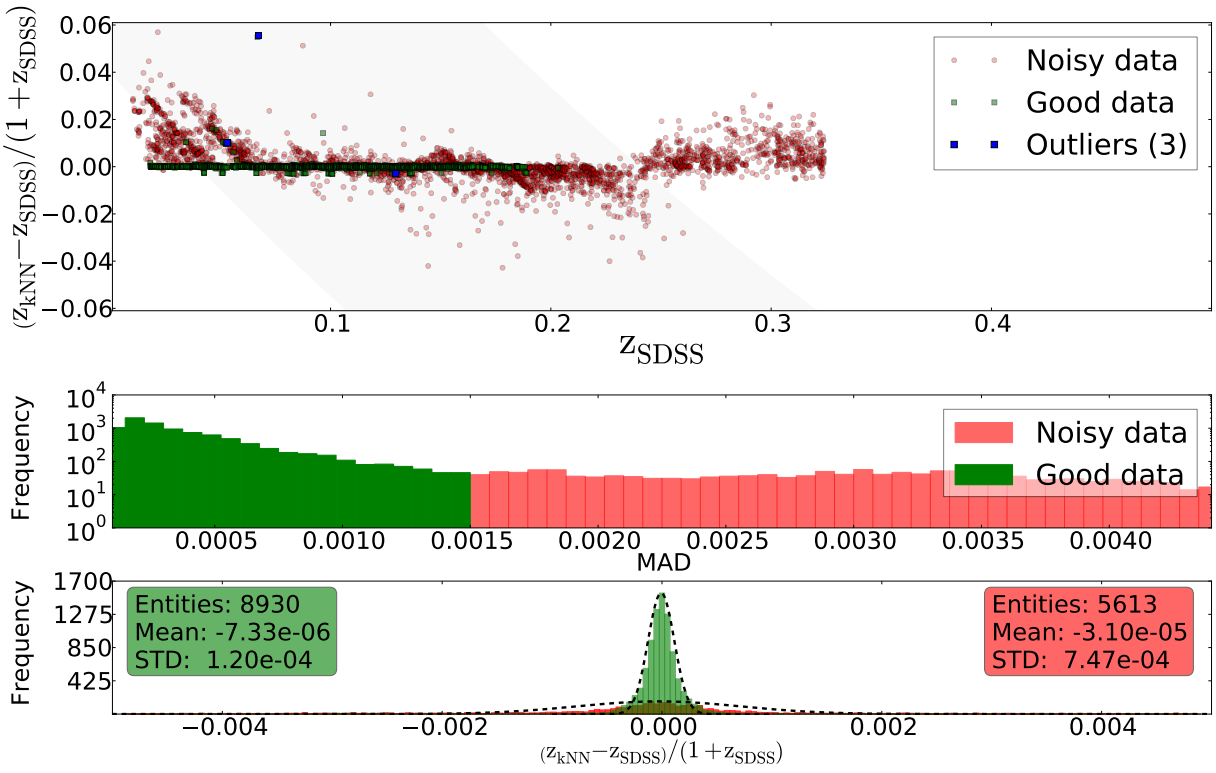


Figure B.8: *Regressional redshift versus spectroscopic one for the emission feature H_α, [NII]*

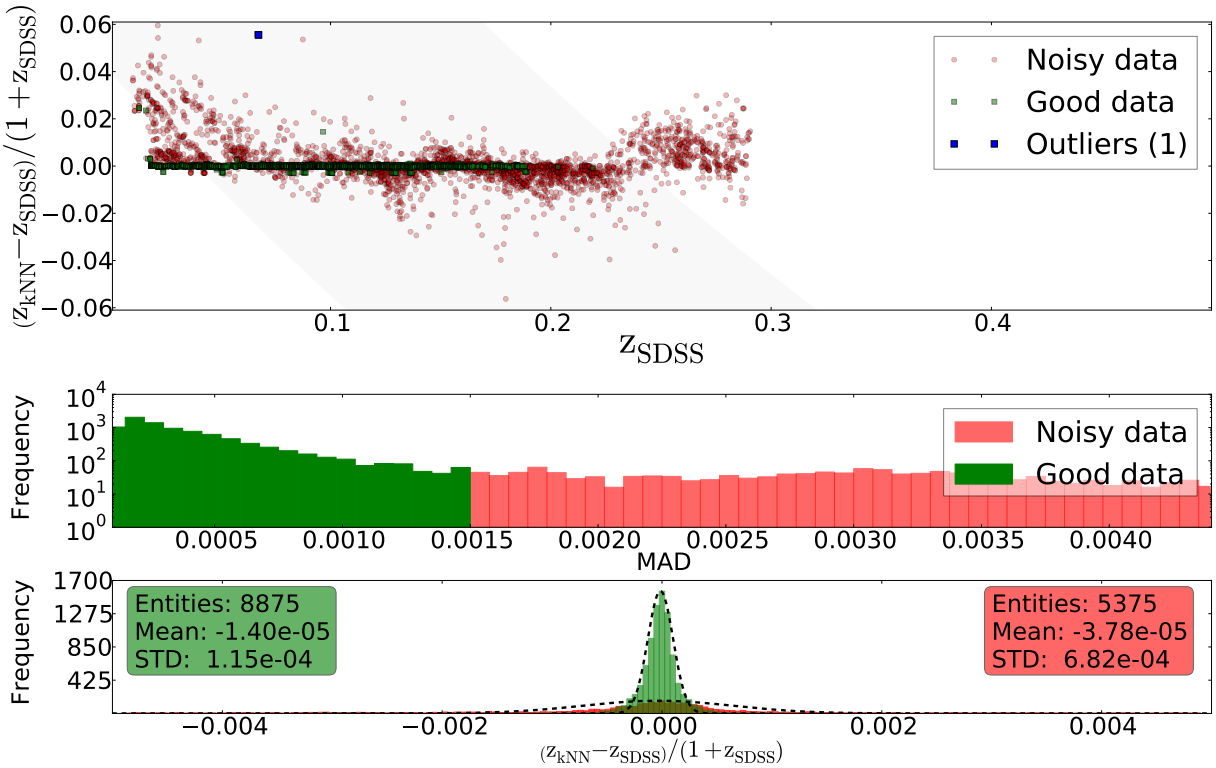


Figure B.9: *Regressional redshift versus spectroscopic one for the emission feature [SII]*

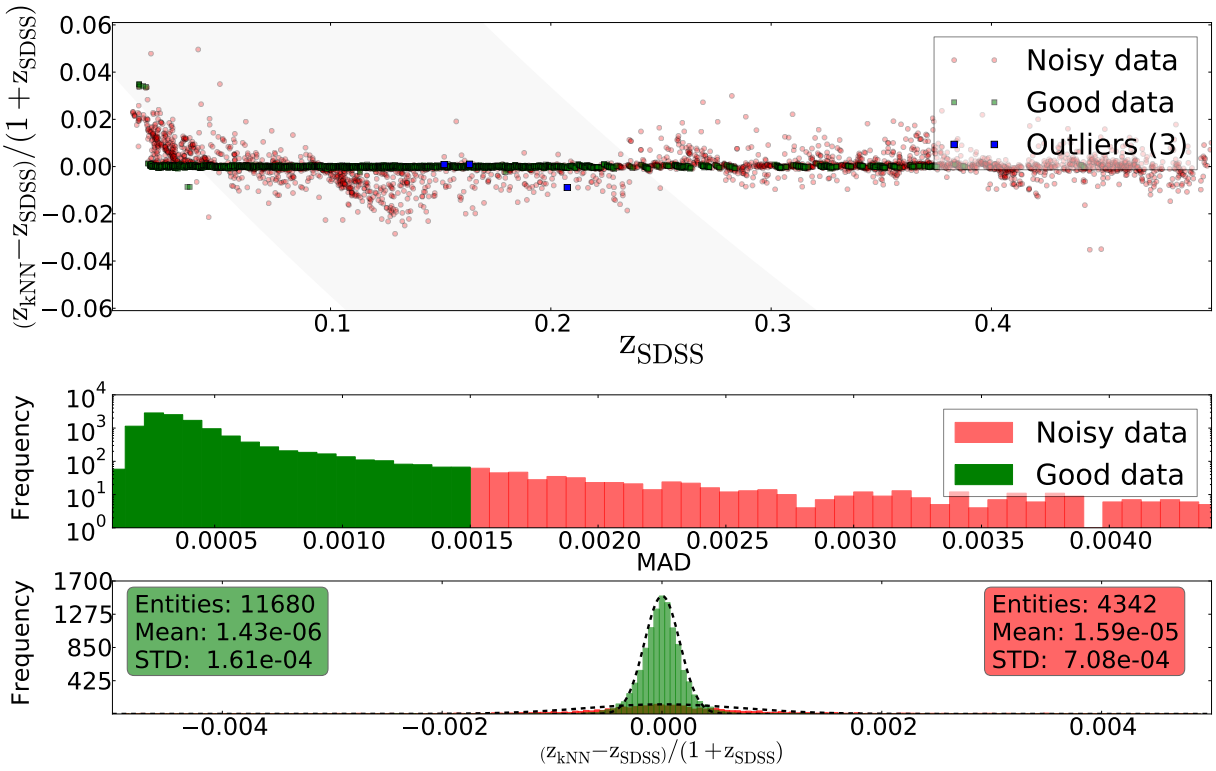


Figure B.10: *Regressional redshift versus spectroscopic one for the absorption feature H+K break*

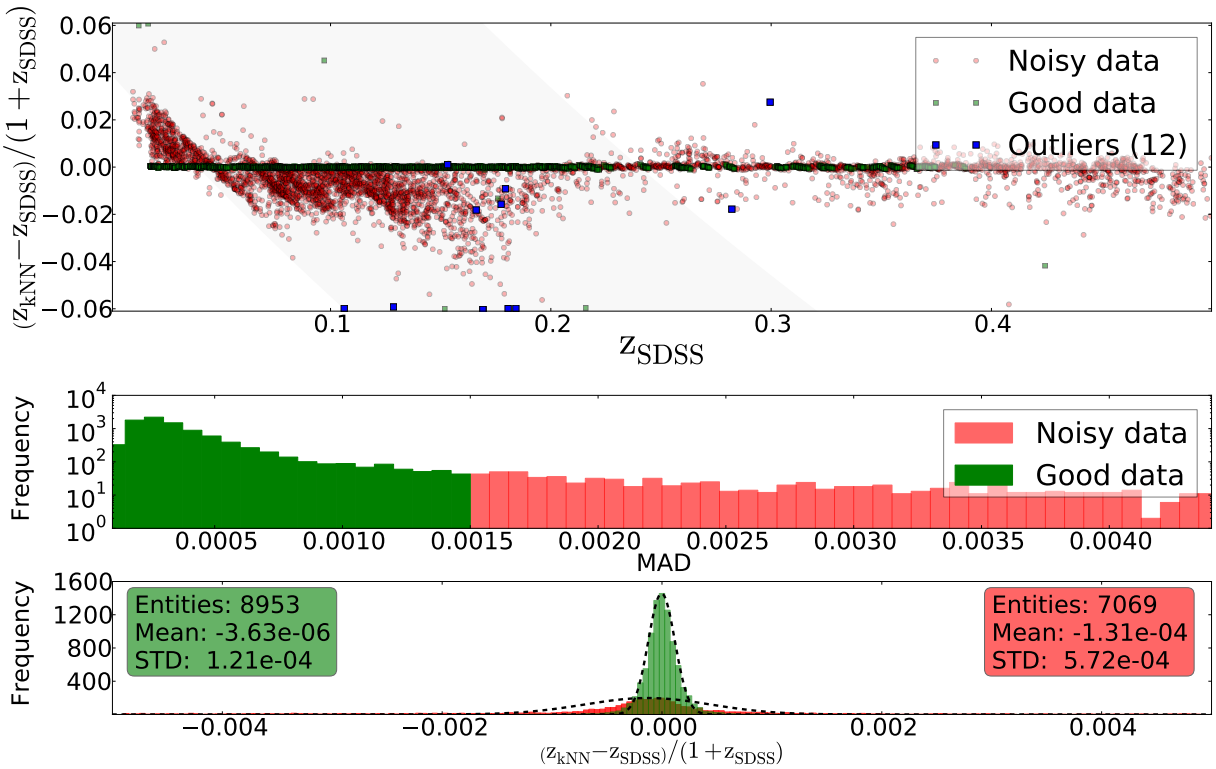


Figure B.11: *Regressional redshift versus spectroscopic one for the absorption feature Mgb*

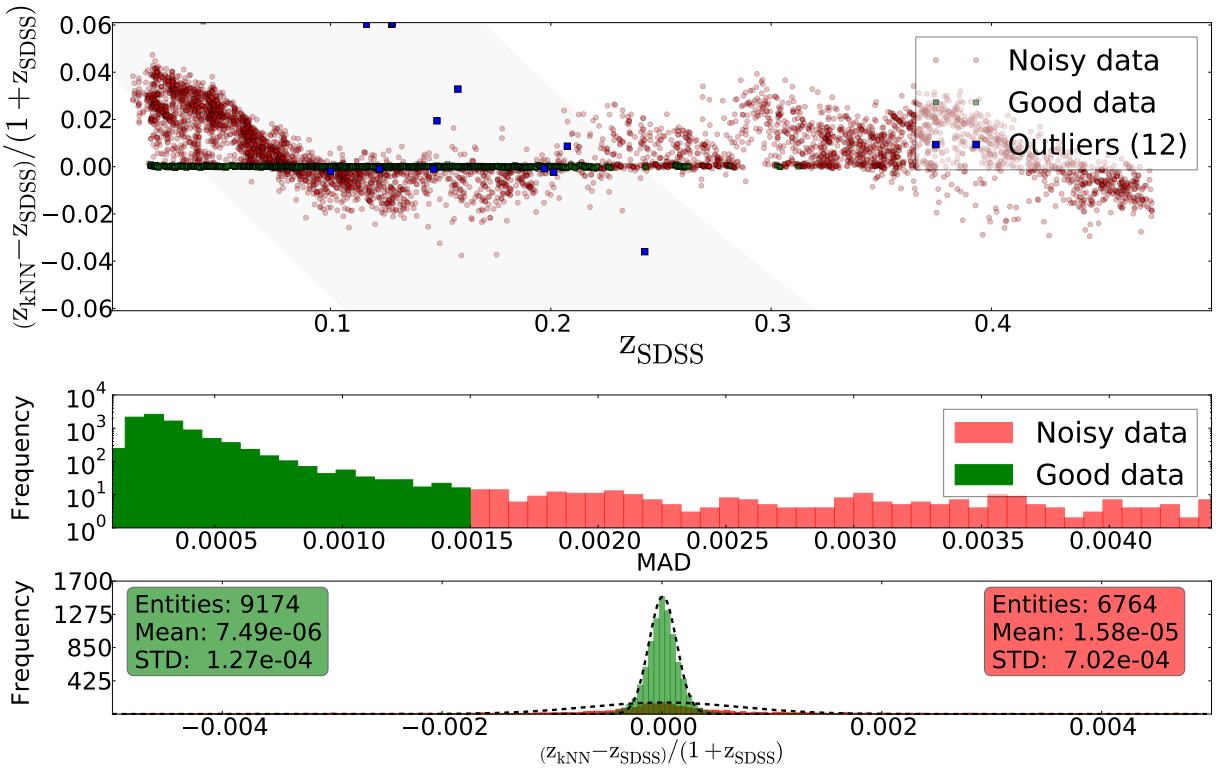


Figure B.12: *Regressional redshift versus spectroscopic one for the absorption feature NaD*

Appendix C

Scripting Tool for the Large Binocular Telescope (LBT)

Part of this PhD thesis was dedicated to the design of a scripting tool for the Large Binocular Telescope (LBT) in Arizona, USA. The scripting tool (SC hereafter) was originally written as a consequence of losses in telescope time due to scripting errors (especially problems with the guide star selection for the AGW, explained below). It is a web tool based on a cgi-form delivering the data from the input form to a Python[18] script, which is then analyzing and checking the data for completeness/correctness using the basic ideas from the scriptchecker tool written by S. Allen [19] (SCT hereafter) and some self-written algorithms (cf. section C.1.3). The requirements for the tool changed as the LBT was approaching the first observations in binocular mode in February 2014, shortly after finishing seeing-limited commissioning of LUCI2 on the right side of the LBT. Apart from delivering error-proven scripts, the scripting tool now had to take into account more severe issues such as synchronized telescope offsets, failures of telescope presets as well as basic functionality of an observation planning tool (which is beyond its original purpose). Thus, the evolution of the scripting tool reflects the developments in software design on instrument side as well as on telescope side (telescope control software; TCS) which were necessary to run the first binocular telescope efficiently. Even though, the tool is self-explanatory on its surface some special features and problematic issues will be explained in some more detail in the sections below.

C.1 Constitution

As already stated the tool consists of three basic layers, namely the input form, the evaluation page and the output (in form of LUCIFER scripts), see also figure C.1.

C.1.1 Form page

The input form consists of the same parts as the out coming script file does, namely

- Archive information
- Telescope setup
- Readout setup
- Instrument setup

[18] https://www.python.org/
[19] http://abell.as.arizona.edu/~lbtsci/Instruments/LUCIFER/Scripts/scriptcheck.html

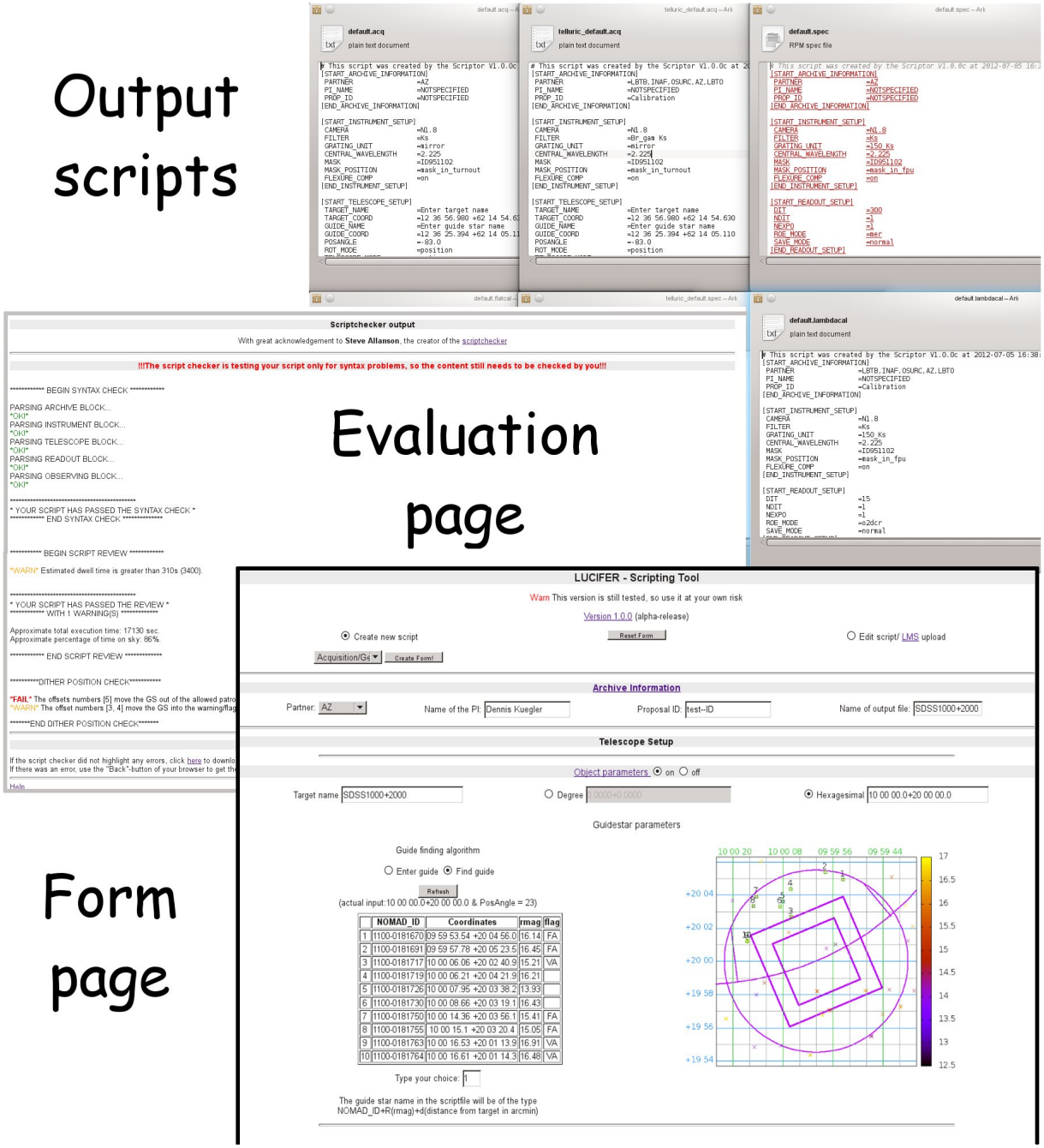


Figure C.1: *Three layers of the SC.*

- Observing setup

plus two further sections:

- Scriptor mode
- Scriptchecker setup

All of these sections can be unselected individually by clicking the respective radio button. This value is then send to the form and is applied to the section after a reload of the page (with the now changed form values). The big advantage of writing all those changes directly into

the form is that all other values are stored immediately, so that a refresh of the page within the program does not reset the input values (internal refresh hereafter). In contrast to this, a refresh of the browser does reset all the input values (external refresh hereafter). The script file sections contain the standard fields as given in the standard scripts, the preset input fields (such as filter wheels and readout mode) are displayed in lists in order to prevent typos from the user. The "mask-position" list, the "refresh" button as well as the "Offset choice" radio button lead automatically to an internal refresh in order to change fields according to the chosen value (unlocking "mask choice", refreshing guide star selection and unlocking upload field for batch files respectively). After entering all the necessary input values (empty fields are ignored) a click of the submit button starts the main script, forwarding the user automatically to the evaluation page.

Scriptor mode

The choice of the scriptor mode fixes the type of script which will be created. After choosing from the list

- Acquisition/Generic
- Spectroscopy
- Imaging
- Calibration
- LMS
- Longslit

the non-necessary parts of the SC are hidden and some standard values are applied (e.g., `readout_mode=mer` for spectroscopy). If a (spectroscopic) calibration script has to be created, the nominal standard exposure times and lamps for the flats and arcs (obtained by testing on the telescope) can be inserted by clicking on the respective button. Furthermore, the respective mode allows the creation of automatized acquisition and calibration files (in Longslit and LMS mode). Despite the guidestar-selection this task might be the most time-saving step when using the scriptor.

Scriptchecker setup

The SCT setup gives the opportunity to let the respective script check with the SCT written by S. Allen. As a sanity check, which after intensive testing might be avoided, all automatically created scripts are processed through the SCT as well, so that systematic errors can be detected. Further details on the SCT are given in the succeeding section.

C.1.2 Evaluation page

The two parts of this page are the output of the SCT setup as well as the link to the final script files. The SCT mainly checks the syntax of the tested script by reading the script line per line and comparing the line to standard line syntax. Some semantic errors, such as the choice of the save and readout mode, are tested as well using basic logic. Another interesting feature of the SCT is that offsets are visualized in an external DS9[20] window, giving the user the possibility to check if the the offset motions move the guide star out of the allowed patrol field (see section

[20] http://hea-www.harvard.edu/RD/ds9/site/Home.html

below). Apart from the latter feature, the outputs of the SCT are directed into a temporary file where the PERL-syntax is translated into HTML-language and is displayed on the evaluation page. As the web-server is not capable of using graphical programs such as DS9, a new dither-position check for the guide star was implemented in the SC. It basically follows the offset steps and gives a warning/error if the an offset moves the guide star into the flagged/forbidden region (for definitions of those regions, see section C.1.3). The scripts (in form of a .zip file in order to force the browser to open the download dialog) can then be downloaded by clicking on the respective link. A sample script for a acquisition script can be seen in listing C.1.

C.1.3 Specific elements

Despite some sanity checks (such as testing the types of input variables) the SC offers more complex and easy-to-use tools which will be explained in the succeeding sections.

Guide star selection

Apart from typos, the selection of the guide star was the major reason for erroneous scripts. The Auto-Guiding and Wavefront sensing unit (AGW hereafter) needs a guide star in order to guide the telescope well and to minimize atmospheric effects causing wavefront distortions. In order to accomplish this, the provided guide star has to be sufficiently bright (13<R<16.5) and has to have a certain position with respect to the detector. The rather unique appearance of this allowed patrol field is caused by the geometry of different optical elements: The different shapes of the constrained regions can be parametrized by the values given in Table C.1, a visualization of the forbidden and flagged regions can be seen in Figure C.2. The guide

Obj	shape	Forbidden region			Flag region			Flag
		center (x,y)[′]	diam./ leng.[′]	end point (x,y)[′]	center (x,y)[′]	diam./ leng.[′]	end point (x,y)[′]	
I	circle	(0,0)	11.0	-	(0,0)	10.0	-	FA
II	square	(0,0)	6.0	-	(0,0)	7.0	-	VA
III	circle	(0,17.15)	34.3	-	(0,17.15)	33.3	-	GA
IVi	line	(-4.50,3.18)	-	(-5.16,0.78)	(-4.08,3.18)	-	(-4.08,0.78)	TA
IVii	line	(4.80,0.72)	-	(3.90,3.90)	(4.08,0.72)	-	(4.08,3.90)	TA

Table C.1: *Parameters for the different patrol field regions for the telescope pointing to RA=DE=PA=0*

star algorithm downloads the respective field from the Naval Observatory Merged Astrometric Dataset (NOMAD[21] hereafter) which is a cross-match of several astrometric catalogs (such as the Hipparcos[22]) and the 2 Micron All Sky Survey (2MASS[23] hereafter). Subsequently, all objects with proper magnitudes are processed through a positional check, in order to exclude objects lying outside the allowed/flagged patrol region. If more than 10 objects are found within the unflagged region, the 10 brightest of them are shown in the HTML table and can be selected by typing their respective table number. If less than 10 objects are unflagged, the table is filled up

[21] http://www.usno.navy.mil/USNO/astrometry/optical-IR-prod/nomad/
[22] http://www.rssd.esa.int/index.php?project=HIPPARCOS catalog
[23] http://www.ipac.caltech.edu/2mass/

```
[START_ARCHIVE_INFORMATION]
 PARTNER =LBTO
 PI_NAME =TestPI
 PROP_ID =test_ID
[END_ARCHIVE_INFORMATION]

[START_INSTRUMENT_SETUP]
 CAMERA =N3.75
 FILTER =J
 GRATING_UNIT =mirror
 #CENTRAL_WAVELENGTH =
 #MASK =
 MASK_POSITION =no_mask_in_use
 FLEXURE_COMP =on
[END_INSTRUMENT_SETUP]

[START_TELESCOPE_SETUP]
 TARGET_NAME =SDSS_140000.0+000000.0
 TARGET_COORD =14 00 00.0 +00 00 00.0
 GUIDE_NAME =NOMAD0900_0234108_R16.97_d3.8
 GUIDE_COORD =13 59 56.66 +00 03 46.9
 POSANGLE =5.0
 ROT_MODE =position
 TELESCOPE_MODE =active
[END_TELESCOPE_SETUP]

[START_READOUT_SETUP]
 DIT =10
 NDIT =1
 NEXPO =1
 ROE_MODE =o2dcr
 SAVE_MODE =normal
[END_READOUT_SETUP]

[START_OBSERVING_SETUP]
 OFFSET_TYPE =relative
 COORD_SYS =DETXY
 OFFSET =10 0
 OFFSET =-20 0
 OFFSET =0 5
 OFFSET =0 10
 ACQUISITION =5 5
[END_OBSERVING_SETUP]
```

Listing C.1: *A sample acquisition script*

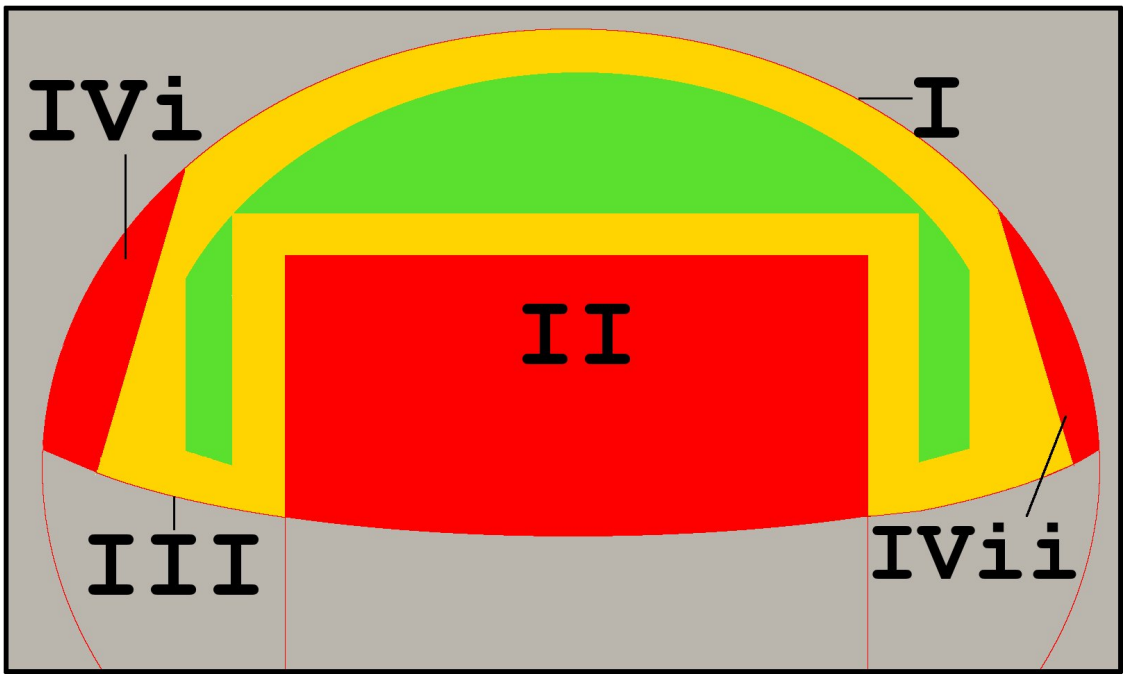


Figure C.2: *The different regions of the patrol field. The forbidden/flagged regions are marked red/orange, the preferred region is marked green.*

with flagged objects until 10 or the maximum of flagged and unflagged objects is reached. By choosing the table number, the guide star is automatically selected for the script file.

Offset batch file

As for a couple of observations (such as exoplanet transits) a huge amount of different offsets is needed, the SC offers the user up to 110 fields which can be used to input their offsets. But as typing such a huge amount of numbers into a web form is an obvious error source a batch file can be uploaded to the web form. If it has the correct format, the page does an internal reload and inputs all the offset values into the respective columns so that it is still possible to apply changes manually.

LMS implementation

Despite the guide star algorithm, the implementation of LMS-file processing is a heart piece of the SC. Lucifer Mask Simulator (LMS) files are produced simultaneously to the multi-object-spectroscopy (MOS) mask which is produced in Tucson/Arizona. This file contains the complete information (such as guide star, filter, grating) for the SC to produce the respective script files. Next to the arcs and flats special spec and acquisition scripts for the telluric are written. The user can select the slit positions into which the telluric should be moved in order to illuminate the complete wavelength range. All in all, the upload of one LMS file yields (if the filter/grating combination is known) 6 clean scripts with which the entire MOS observation can be accomplished. Further fields are added in LMS mode in order to guarantee that no manual user-input is needed, e.g., for choosing filters for the target or telluric acquisition.

C.2 Upgrade to binocular observations

With the completion and installation of the second LUCIFER instrument, the possibility of performing binocular observations arose. Binocular observations were the key motivation for building the LBT. Binocular thereby denotes the parallel (synchronized) observations with the two different mirrors using the corresponding (homogeneous) or a different (inhomogeneous) instrument on the other side. The hope of this observation type is to gain a huge flexibility in observations (e.g., simultaneous spectra from the near-IR to the optical) and eventually, also save valuable and costly observation time.

In order to run a telescope efficiently in binocular mode, preparation tools have to be made available to the community that are reliable, and easy to understand. Consequently, with the new instrument also a new format of the scripts was introduced, which is based on the *XML* format. An example of the new script type is shown in Listing C.2

The *XML* format is a machine-readable format and thus no conversion between plain ASCII files and the *XML* has to be done. Additionally, constraints on the entered values can be set in *XML* and thereby effectively avoiding typos or invalid values.

In order to prepare the SC for binary observations, several new things had to be added. Apart from introducing new respective options on the form page, an obvious visualization tool for understanding the losses in time due to the use of the binocular mode had to be provided. Several other constraints arise from the pointing limit, which indicates the maximum angle the pointings of the two mirrors can differ, without violating the hardware limits. Eventually, the SC has to be able to arrange the observation items in the order that was intended by the user.

The installation of LUCIFER 2 enabled also the use of adaptive optics, as it is the first instrument with the high-resolution camera N30. The experience with the adaptive optics system is still very limited, so that only approximated times (e.g., for closing the AO-loop) for the observation in diffraction-limited mode are available. Consequently, the overall execution times of the scripts are highly uncertain and more development is needed to also account for potentially fatal events, like a break of the AO-loop.

```
<observationProgram>
<label>Imaging-script by NOTSPECIFIED with LUCI2,LUCI1</label>
<partner>LBTB</partner>
<pi>NOTSPECIFIED</pi>
<propID>NOTSPECIFIED</propID>
<observationBlocks>
<observationBlock idBlock="1">
        <observationItems>
                <observationItem idItem="1">
                        <itemType>REGULAR</itemType>
                        <useBinocularMode>true</useBinocularMode>
                        <synchronizeInstruments>true</synchronizeInstruments>
                        <telescope>
                                <mount objectName="Standard Star A" ra="300.0" dec="+10.0" epoch="2000.0" />
                                <mirrors>
                                <mirror usageType="1">
                                        <rotationAngle angleType="POSITION_ANGLE">20</rotationAngle>
                                        <offset offsetType="0" absoluteOffset="true" equatorialCoordinate="false">
                                                <coord>7.5</coord>
                                                <coord>-0.2</coord>
                                        </offset>
                                <guideStarsAuto>false</guideStarsAuto>
                                <guideStars>
                                        <guideStar objectName="!N1000-0563247_R13.48_d0" ra="300.051" dec="10.066" epoch="2000.0" />
                                </guideStars>
                                <useActiveOptic>true</useActiveOptic>
                                <useAdaptiveOptic>false</useAdaptiveOptic>
                                </mirror>
                                <mirror usageType="2">
                                        <rotationAngle angleType="POSITION_ANGLE">20</rotationAngle>
                                        <offset offsetType="0" absoluteOffset="true" equatorialCoordinate="false">
                                                <coord>4.6</coord>
                                                <coord>7.4</coord>
                                        </offset>
                                <guideStarsAuto>false</guideStarsAuto>
                                <guideStars>
                                        <guideStar objectName="!N1000-0563247_R13.48_d0" ra="300.051" dec="10.066" epoch="2000.0" />
                                </guideStars>
                                <useActiveOptic>true</useActiveOptic>
                                <useAdaptiveOptic>false</useAdaptiveOptic>
                                </mirror>
                                </mirrors>
                        </telescope>
                        <instruments>
                                <instrument usageType="1">
                                        -<lamps></lamps>
                                        <lampsTime></lampsTime>
                                        <calibrationModeStatus>0</calibrationModeStatus>
                                        <cameraWheelPosition>1</cameraWheelPosition>
                                        <filters>
                                                <filter>2</filter>
                                                <filter>4</filter>
                                        </filters>
                                        <gratingPosition>1</gratingPosition>
                                        <maskPosition>2</maskPosition>
                                        <flexureCompState>true</flexureCompState>
                                </instrument>
                                <instrument usageType="2">
                                        -<lamps></lamps>
                                        <lampsTime></lampsTime>
                                        <calibrationModeStatus>0</calibrationModeStatus>
                                        <cameraWheelPosition>1</cameraWheelPosition>
                                        <filters>
                                                <filter>2</filter>
                                                <filter>2</filter>
                                        </filters>
                                        <gratingPosition>1</gratingPosition>
                                        <maskPosition>2</maskPosition>
                                        <flexureCompState>true</flexureCompState>
                                </instrument>
                        </instruments>
                        <detectors>
                                <detector usageType="1">
                                        <dit>20</dit>
                                        <ndit>1</ndit>
                                        <numberOfReads>2</numberOfReads>
                                        <nexp>1</nexp>
                                        <roeMode>LIR</roeMode>
                                        <saveMode>NORMAL</saveMode>
                                        <frameType>SCIENCE</frameType>
                                </detector>
                                <detector usageType="2">
                                        <dit>30</dit>
                                        <ndit>1</ndit>
                                        <numberOfReads>2</numberOfReads>
                                        <nexp>2</nexp>
                                        <roeMode>LIR</roeMode>
                                        <saveMode>NORMAL</saveMode>
                                        <frameType>SCIENCE</frameType>
                                </detector>
                        </detectors>
                </observationItem>
                -<observationItem idItem="2"></observationItems>
        </observationItems>
</observationBlock>
</observationBlocks>
</observationProgram>
```

Listing C.2: *A sample binocular imaging script*

Acknowledgements

I would like to thank my family for their emotional and financial support over all the years and for encouraging me to pursue my studies.

I would also like to thank my supervisor Jochen Heidt who convinced me to start this PhD thesis and brought me into the LUCI project.

I thank Kai for introducing me into the wonderful world of data-driven science.

I am very grateful for having Nikos as my colleague, without him I would not have been able to finish my thesis in this time.

Christian Fendt, for his efforts as an IMPRS supverisor.

My proofreaders Nikos, Dorothea, Volker, Kai & Benny.

All the ones, I have forgotten.

Eidesstattliche Erklärung

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

Heidelberg, den 24. September 2015

_____

Sven Dennis Kügler