UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

PhD Thesis

Recycling on a Cosmic Scale

Extracting New Information from Old Data Sets

Author:
Kristoffer Stensbo-Smidt

Supervisors:
Kim Steenstrup Pedersen
Christian Igel

Submitted
10th November 2016
to the PhD School of The Faculty of Science, University of Copenhagen

Abstract

Astronomy and astrophysics are entering a data-rich era. Large surveys have, quite literally, seen the light in the past decade, with more and larger telescopes to follow in the coming years. Data is now so abundant that making use of all the information is a difficult tasks. This thesis sets out from the assumption that there is more to gain from available data sets – new information from old data. Three contributions in this direction are considered.

Firstly, a novel texture descriptor for parametrising galaxy morphology is presented. It uses the shape index and curvedness of local regions in images of galaxies and condenses information about the local structure to a single value. It is argued that this value can be interpreted as indicating regions of morphological interest, for example regions of newly formed stars, of gas and dust, spiral arms etc. The descriptor is shown to extract information about a galaxy's specific star formation rate from its images that the usual spectra energy distribution (SED) fitting misses.

Secondly, a method to evaluate the information content of various features for a given task is introduced. Selecting the right features, for example colours or magnitudes, for a specific task can be difficult and often relies on which have been used traditionally. With current and future surveys giving researchers access to hundreds of features, it is time to challenge old assumptions on which to use. A completely general method for feature selection is introduced and shown to increase accuracy of both redshift and specific star formation estimations.

Thirdly, the problem of quality assessment of quasar candidates is considered. Detection pipelines searching the sky for quasars produce thousands of candidates, many of which can be discarded with simple checks. The rest, however, cannot, and images of these candidates must be manually inspected and evaluated. Still, more than 90% of these can be false positives, wasting precious time for researchers and forcing a limitation of the scopes of the detection pipelines. A set of features based on image analysis is presented and shown to be able to detect the most common situations of false positive quasar candidates. Incorporation of the derived features into a machine learning frameworks is reviewed and future directions are discussed.

Acknowledgements

A great many people played keys parts in making this thesis reality. First and foremost, I thank my supervisors, Kim Steenstrup Pedersen and Christian Igel, for advising me throughout the past three years. Your guidance and our fruitful discussions have been invaluable.

Many thanks goes to the kind people at the Imperial Centre for Inference and Cosmology, Imperial College London, for making me part of your group for six months. In particular, I would like to thank Daniel Mortlock, Stephen Warren, Alan Heavens, Joshua Greenslade, Ciarán Conneely, Claude Schmit, Charlotte Norris, Rhys Barnett, and Justin Alsing for being kind, open and welcoming, and to the London Farmers' Market for making Tuesday's lunch much more exciting.

I would also like to thank the open and welcoming environment at the Image Section and all my friendly colleagues here. My office mates, in particular, have made the life as a PhD student much more entertaining than it is rumoured to be. Due to the dynamic nature of the Department of Computer Science, there have been too many office mates to name you all, but Jan Kremer, Niels Dalum Hansen, and Oswin Krause have made the time in the office most enjoyable and interesting. A sincere thank you goes to the coffee machine for helping me through long days, nights, and weekends.

Last, but not least, I thank my family, Mathias, Lise and Søren, for incredible support throughout the years. The biggest thank you goes to my better half and seal, Tina Ibsen, whose wisdom, hugs, love and proofreading made this thesis possible and even the most unpleasant and stressful times bearable.

This entire PhD has been made possible by help from Henrik Brink and by funding from the Danish Council for Independent Research | Natural Sciences for the project *SkyML – Surveying the sky using machine learning*. My stay in London would not have been possible without financial support from Oticon Fonden, Knud Højgaards Fond, and Aage og Johanne Louis-Hansens Fond.

Contents

List of Figures

1

Preface

Astronomy is at this very moment undergoing a paradigm shift. Transitioning from a time of scarce data to a time of data so plentiful that it takes dedicated efforts just to store and access it. As a consequence of this, the field of *astroinformatics* or *astrostatistics* is evolving – an interdisciplinary field of astronomers, statisticians, computer scientists and data scientists. Extensive surveys, such as Sloan Digital Sky Survey (SDSS), have made it possible to do science in a way never before seen in astronomy, giving researchers access to information about a billion objects in the sky at the click of a mouse. And SDSS was only the beginning.

Many new astronomical observatories are being planned or built as you read this. The Large Synoptic Survey Telescope (LSST), scheduled to be fully operational in 2022, will produce about 30 terabytes of images per night, which need to be analysed in near real-time to detect fast changing sources, so-called transients. Another telescope, the Square Kilometre Array (SKA), expected to be operational by the end of the 2020s, will produce a massive 1 exabyte of raw data per night. Needless to say, manual processing is out of the question.

Advanced methods from machine learning and computer vision have slowly entered astronomical research in the past couple of decades, but there is still much to do. There are many open questions in astronomy that could benefit from the advanced statistical methods available in the computer science field, and there are many interesting problems in astronomy that could spark new ideas and approaches in both computer science and statistics communities.

This thesis has been submitted for the degree of PhD, and was conducted at the Faculty of Science, University of Copenhagen. The PhD was funded by The Danish Council for Independent Research | Natural Sciences through the project *SkyML – Surveying the sky using machine learning*. A major objective has been to show the relevance of interdisciplinary research in today's academic environment. By combining astrophysics, statistics, and computer science, we aim to show that interdisciplinary research can benefit all involved parties.

A core hypothesis of this thesis is the idea that there is much more to be learnt from already available data sets, and that advanced statistical methods, such as machine learning and computer vision, can help uncover this information. This is investigated in three different projects covering texture analysis of galaxies, feature selection for redshift estimation, and quality assessment of quasar candidates.

The structure of the thesis is as follows. Chapter 2 introduces basic astronomy and astrophysics required to understand the astronomical motivation for the projects. Chapter 3 covers background and additional information, which can help understand the work done in the projects, as well as discusses some of the specific choices we have made. Chapter 4 provides a summary of the results and conclusions for each of the included papers, as well

as the work in progress. Finally, chapter 5 summarises the projects, offers perspectives on the experience gained from them, and discusses some future directions for continued work.

The following papers have been produced as part of this PhD. This thesis builds on the first three papers, which can be found in part II.

- J. Kremer, K. Stensbo-Smidt, F. Gieseke, K. Steenstrup Pedersen, and C. Igel. Big universe, big data: machine learning and image analysis for astronomy. *IEEE Intelligent Systems*. Accepted for publication, September 2016.
- K. Steenstrup Pedersen, K. Stensbo-Smidt, A. Zirm, and C. Igel. Shape index descriptors applied to texture-based galaxy analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2440–2447, 2013.
- K. Stensbo-Smidt, F. Gieseke, C. Igel, A. Zirm, and K. Steenstrup Pedersen. Sacrificing information for the greater good: how to select photometric bands for optimal accuracy. *Monthly Notices of the Royal Astronomical Society*, 464(3):2577–2596, 2017.
- K. Stensbo-Smidt, C. Igel, A. Zirm, and K. Steenstrup Pedersen. Nearest Neighbour Regression Outperforms Model-based Prediction of Specific Star Formation Rate. In *Proc. IEEE Int. Conf. Big Data*, pages 141–144, 2013.

Part I

Background

2

A brief introduction to astronomy

For centuries, astronomy has suffered from small sample sizes as acquiring data requires access to large and expensive telescopes. Furthermore, the data acquired has often been of low quality due to the vast amount of effects influencing data collection: faint objects, atmospheric disturbances, light pollution, mirror defects, noise and variabilities in electronic devices, etc. This is changing now.

Astronomy has only in recent decades begun to acquire data in large enough amounts, and of good enough qualities, that precision measurements of the fundamental parameters governing galaxies and the Universe can be done. A wealth of new knowledge has emerged from this, for instance the discovery of dark energy, which makes the expansion of the Universe accelerate and accounts for 70% of the content in the Universe (Riess et al., 1998; Perlmutter et al., 1999). We have also learnt much about the evolution of galaxies, for instance that there appears to be (at least) two fundamental classes of galaxies, namely inactive ellipticals and active spirals.

Every new insight, however, has led to many new questions. For instance, the transition from active to inactive galaxy appears to be incredibly fast, and the mechanisms responsible for this quenching are still not known.

Modern surveys, such as the Sloan Digital Sky Survey (SDSS, York et al., 2000) and the UKIRT Infrared Deep Sky Survey (UKIDSS, Lawrence et al., 2007), have provided us with incredible amounts of data. For the first time in the history of astronomy, doing large-scale statistics is possible, and upcoming surveys will only add to this. In fact, even today's survey data prove challenging for researchers to handle; a challenge that will only become larger. The data quantities today require automated methods for extracting useful information. The available methods, however, still have many shortcomings, and data appears increasingly difficult to handle.

This thesis investigates a few of the problems faced by astronomers today and approaches them from a machine learning and computer vision perspective. The main problems attacked are redshift estimation, star formation rate estimation and detection of quasar candidates. Below is a short introduction to these problems. For an introduction to the intersection of machine learning and astronomy, see the review by Kremer et al. (2016), chapter 6.

2.1 Redshift estimation

One hundred years ago, it was believed that the Universe was static. All galaxies were thought to be at rest in space, exactly where they had always been and where they would always be. In 1929, however, the astronomer Edwin Hubble showed that the light from galaxies was *redshifted*, a phenomenon caused by the Doppler effect. This effect causes a

reddening of the light, when the object moves away from us. Correspondingly, an object moving towards us is blueshifted.

Redshift, z, is measured by comparing the wavelength of the observed light and that of the emitted light,

$$z = \frac{\lambda_{\text{obs}} - \lambda_{\text{emit}}}{\lambda_{\text{emit}}} . \tag{2.1}$$

In practice, one can do this by looking at how the spectral 'fingerprints' of atoms in the galaxies have been shifted compared to the same atoms here on Earth.

The fact that the light from a galaxy is redshifted means that it is moving away from us. Hubble (1929) demonstrated that the redshift is correlated with distance, and this seemingly innocent observation has had major implications for astronomy. First of all, it showed that the Universe is not static, but expands in all directions. Secondly, measuring distances in the Universe is notoriously difficult, but suddenly there was a way of doing this. In fact, since light has a finite speed, looking further into the Universe means looking back in time. Thus, redshift can not only used to specify a galaxy's distance to us, but also the time at which the light was emitted. That is, the higher the redshift of a galaxy, the older a snapshot of the Universe we see.

Redshift is due to the Universe expanding, 'stretching' the light as it propagates through space. The scale factor a of the Universe and the redshift are inversely related,

$$a = \frac{1}{1+z} . \tag{2.2}$$

Today $a = 1$, and at the time of the Big Bang, $a = 0$. Thus, at redshift $z = 1$ the Universe was half the size it is today. Using models of how the Universe has evolved allows us to translate redshift, or, equivalently, size of the Universe, to the age of the Universe at the given redshift.

The fact that redshift and distance are correlated made it possible to construct large-scale maps of the Universe, showing the distribution of galaxies in large parts of the observable Universe through time, in turn allowing for inference of some of the most fundamental parameters of the Universe.

Getting accurate redshifts is still a major issue, as detailed by Calcino and Davis (2016) – getting accurate redshifts for a large number of galaxies is an even bigger one. For accurate redshifts, one needs a high-resolution spectrum of the light from a galaxy, which is expensive and time-consuming to obtain. Getting images of galaxies in broad-band filters, on the other hand, is comparatively cheap and fast. They are, however, also of much lower quality, since a spectrum containing thousands of individual measurements is reduced to only a handful. This is illustrated in Fig. 2.1, which shows a spectrum obtained from the SDSS spectrograph together with the broadband filters used for the camera.

The cost of obtaining a spectrum means that they are vastly outnumbered by images of galaxies. This is exemplified by the SDSS data base, which contains images of about 200 million galaxies, but only spectra of about one million of those. Images are also able to see much fainter objects than spectra, meaning that we can look deeper into the Universe and, equivalently, further back in time and study the evolution of galaxies.

Thus, obtaining accurate redshifts from imaging data alone can increase our knowledge of the universe substantially, a task that machine learning is well-suited for and is one of the focus areas of Stensbo-Smidt et al. (2017).

2.2 Star formation rate in galaxies

Quantifying a galaxy's evolutionary state is not straightforward, and there are many different measures applied for this. One is the star formation rate (SFR), which measures the number of stars being formed in the galaxy per year – how active the galaxy is.

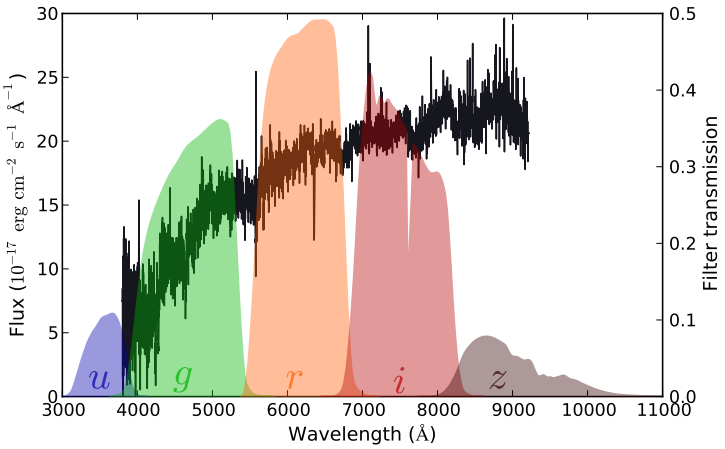


Figure 2.1 An example spectrum of a galaxy from the SDSS database (black curve) overlaid by the five broadband filters of SDSS (Fukugita et al., 1996). In the spectrum, the intensity of the light has been measured at thousands of wavelengths, which are reduced to five *magnitudes*, quantities of integrated light, by the broadband filters, thus significantly reducing the amount of spectral information. As the information from the broadband filters comes in the form of images, however, additional information can be extracted from the spatial structure.

There are many interesting aspects of the SFR. For example, it appears like galaxies in the Universe overall became more and more active until $z \approx 2$, at which point the overall activity began to decrease and has been decreasing since. We still don't know what has caused this effect, which seems to be affecting the Universe as a whole.

From the data gathered by SDSS we have learnt that there exist two major classes of galaxies: the red and dead galaxies and the blue and active galaxies (Kauffmann et al., 2003). It is believed that the active galaxies will eventually die and join the red and dead group of galaxies, but the transition seems to be happening incredibly fast (on cosmological timescales). So fast, in fact, that astronomers have yet to find an explanation for what is 'killing' them.

Achieving good statistics on the SFR of galaxies in large regions of the Universe is paramount to understand what is going on. As with redshift, the SFR of a galaxy is usually estimated from a spectrum of its light, which is expensive to acquire. If it could be reliably estimated from images of the galaxy, we could vastly improve on the statistics.

Image analysis seems to be a particular promising path for SFR estimation, as star formation should change the appearance (morphology) of a galaxy: newly formed stars are bright and blue, and large regions of obscuring dust show where new stars are likely to be forming. This information is not currently used for SFR estimation, so astronomy could potentially gain a lot from an image analysis perspective. This is exactly what was done by Steenstrup Pedersen et al. (2013).

2.3 Detecting distant quasars

There are many subgroups of galaxies within the two major ones described above. One class of particularly active galaxies are called *quasars*. A quasar is in fact just a tiny region around a supermassive black hole in the centre of the galaxy, emitting many times more light than the entire galaxy they reside in.

Quasars are interesting to study exactly because of their incredibly brightness. In fact, they can be seen across most of the observable Universe. Thus many quasars are incredibly old – the oldest and most distant one known is seen when the Universe was a mere

800 million years old, less than 6% of the current age (Mortlock et al., 2011). Studying the light from a quasar can tell us what galaxies consisted of and thus what the early Universe looked like. Furthermore, if the light interacts with matter along the way to us, for instance a diffuse gas cloud, this will leave a 'fingerprint' in the spectrum of the light. We can therefore not only learn about the Universe at the time of the quasar, but how it has changed through time.

Observing distant quasars can provide key pieces of the puzzle of understanding the evolution of the Universe. First, however, one needs to locate the quasars, which is not an easy task. They appear as tiny point sources on the sky, and the most distant ones are only visible in the infrared part of the spectrum.

Astronomers employ detection pipelines scanning the sky for these distant quasars, but the pipelines pick out many false positives. Often times they get confused by nearby objects or spurious effects, needing manual assessment of thousands of images. Image analysis and machine learning can automate a great deal of the work currently done by astronomers, making it possible to increase the scope of the search tremendously, thus increasing the chance of detecting these valuable objects.

Section 3.3 describes ongoing work aiming at detecting a quasar at $z > 8$. At such high redshifts only very few quasars are expected to be visible, so searching for them requires scanning large regions of the sky. This will lead to many thousands of false positive candidates, necessitating automated image analysis software that can detect and discard the majority of these. Examples of false positive candidates are shown, and our progress on detecting these is discussed.

3

Methodology

This chapter presents background and additional information aiding the understanding of the papers included in this thesis. Section 3.1 introduces the background and derivations of the texture descriptors proposed by Steenstrup Pedersen et al. (2013). Section 3.2 reviews some standard methods for feature selections and discusses the reasons for choosing the one used by Stensbo-Smidt et al. (2017). Section 3.3 discusses the background and preliminary results of the task of quality checking images of quasar candidates. This work is still unpublished.

3.1 Morphology of galaxies

It has been known since the days of Edwin Hubble that galaxies can be divided into classes based on their visual appearance, the most widely used classification scheme being defined by Hubble himself (Hubble, 1926). The Hubble sequence extends from featureless elliptical galaxies to structure-rich spiral galaxies. Some spirals feature a bar, leading to the Hubble sequence often being referred to as the 'Hubble tuning fork', see Fig. 3.1. Countless works have later shown how this classification correlates with many intrinsic parameters, such as their environment and evolutionary state (e.g. Skibba et al., 2009). In particular, morphology can be indicative of the processes driving a galaxy's current star formation. For instance, multiple nuclei and irregular features hint at a recent merger. Determining the morphology of galaxies, in particular at high redshifts, can therefore help us understand why star formation has varied through time and how the Universe ended up looking like it does today.

While the Hubble sequence continues to be one of the most valuable morphological classification schemes, it requires manual inspection to determine the Hubble class of a galaxy. Not only is such an approach incredibly demanding in terms of labour, it is also very subjective. Recently, the highly successful citizen science effort Galaxy Zoo (Lintott et al., 2008) has provided the scientific world with a most valuable data set of manual classifications of close to a million galaxies. Still, it is of great interest to be able to quantify the morphology of a galaxy in an objective way, and multiple measures have been used throughout time.

Some of the most commonly used morphological measures are

- the apparent ellipticity (Hubble, 1926),
- the Sérsic index, measuring the shape of the light profile (Sérsic, 1963),
- the Gini coefficient, a rank-ordered cumulative distribution function of a galaxy's pixel intensities (Abraham et al., 2003),

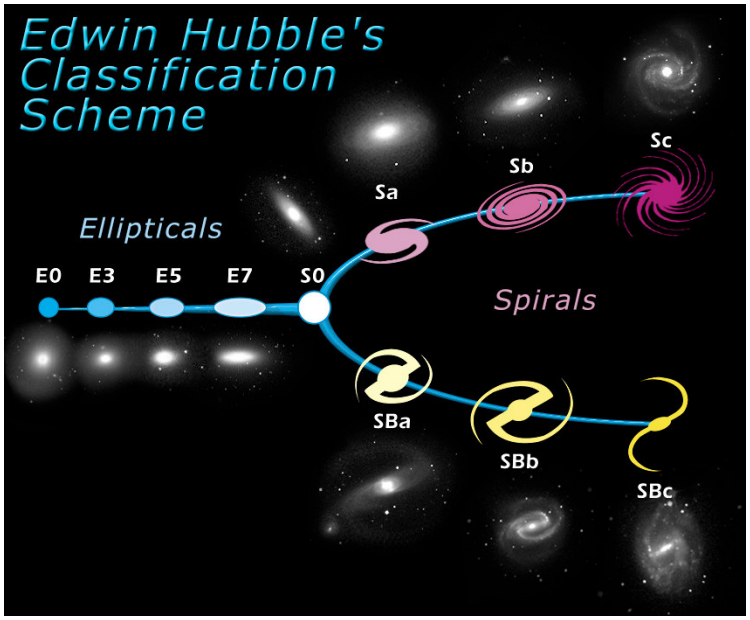


Figure 3.1 The Hubble classification scheme, often referred to as the 'Hubble tuning fork'. Edwin Hubble also defined a third class, 'irregular' galaxies, which are characterised by a chaotic structure. Credit: NASA & ESA.

- the M_{20}, the second-order moment of the brightest 20% galaxy pixels (Lotz et al., 2004),
- the concentration (C), measuring the ratio of light in an inner aperture to that in an outer (Abraham et al., 1994, 1996; Bershady et al., 2000),
- the rotational asymmetry (A) (Schade et al., 1995),
- and smoothness or clumpiness (S), measuring the amount of small-scale structure in a galaxy (Conselice, 2003).

The latter three methods are often combined to form the *CAS* morphological system (Conselice, 2003).

Many of the standard morphological measures focus on the general shape of the galaxy, ignoring the finer structure within it. An exception is the clumpiness, which is based on subtracting the background and a smoothed version of the image from itself (Conselice, 2003),

$$S = 10 \times \sum_{x,y=1,1}^{N,N} \frac{(I_{x,y} - I^{\sigma}_{x,y}) - B_{x,y}}{I_{x,y}}, \tag{3.1}$$

where $I_{x,y}$ is a pixel at (x, y) in the $N \times N$ pixel image, $I^{\sigma}_{x,y}$ is a pixel in an image smoothed by a Gaussian filter at scale σ, and $B_{x,y}$ is a background pixel. Subtracting the smoothed image from itself will reveal small-scale fluctuations. In fact, the clumpiness is a special case of the *difference of Gaussians* method from image analysis, which is used as a blob detector. For blob detection one would typically search different smoothing scales to reveal differently sized structures. The smoothing scale for the clumpiness, however, is usually predetermined as some fraction of the Petrosian radius, a redshift-independent measure of the radius of a galaxy (Petrosian, 1976).

Steenstrup Pedersen et al. (2013), chapter 7, approached the problem of describing galaxy morphology from a texture description point of view. By introducing texture descriptors for first and second order differential structure we were able to capture information about the star formation rate of galaxies not available through the measured magnitudes.

Texture descriptors

There are many different approaches to describing texture. Well-known approaches to texture classification include local binary patterns (Ojala et al., 2002) and textons computed from responses from filter banks (Varma and Zisserman, 2005). Other texture descriptors, such as Basic Image Features (BIF, Griffin and Lillholm, 2007), and patch-based descriptors, such as Scale Invariant Feature Transform (SIFT, Lowe, 2004), Histograms of Oriented Gradients (HoG, Dalal and Triggs, 2005) and DAISY (Tola et al., 2010), approach the problem using differential geometry, for instance by using first order differential information. We take the same differential geometry route, adding also second order information to our descriptor.

As we will be working with differential structure, we need to compute image derivatives. We also want to capture texture at different scales, thus we need to use a multi-scale representation of the images, a so-called *scale space*. Working in scale space means that the computations are less prone to noise in the image, and we furthermore get the image derivatives almost for free.

For an image $I : \Omega \to \mathbb{R},\ \Omega \in \mathbb{R}^2$, we define the *scale-space representation* in terms of a convolution with a Gaussian filter G as $L(x, y; \sigma) = (I * G)(x, y; \sigma)$, where $\sigma > 0$ indicates the scale of the filter:

$$G(x, y; \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) . \tag{3.2}$$

Since the differential and convolution operators commute, any derivative of an image can easily be computed by convolving the image with the corresponding derivative of a Gaussian filter:

$$L_{x^n y^m}(x, y; \sigma) = \left(I * \frac{\partial^{(n+m)} G}{\partial x^n \partial y^m}\right)(x, y; \sigma) . \tag{3.3}$$

To make images comparable across scales, we need to scale normalise the Gaussian filter. which is done by multiplying the derivative with the scale for each order of differentiation (Lindeberg, 1994). The scale normalised image derivative can thus be written as:

$$\sigma^{(n+m)} L_{x^n y^m}(x, y; \sigma) = \left(\sigma^{(n+m)} I * \frac{\partial^{(n+m)} G}{\partial x^n \partial y^m}\right)(x, y; \sigma) . \tag{3.4}$$

Thus, first order structure (gradients) can be written as

$$\mathbf{g} = \sigma \nabla I = \sigma \left(\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}\right)^\mathsf{T} \equiv \sigma (L_x, L_y)^\mathsf{T} , \tag{3.5}$$

whereas the second order structure (the Hessian) can be written as

$$\mathbf{H} = \sigma^2 \nabla^2 I = \sigma^2 \begin{bmatrix} \frac{\partial^2 I}{\partial x^2} & \frac{\partial^2 I}{\partial x \partial y} \\ \frac{\partial^2 I}{\partial y \partial x} & \frac{\partial^2 I}{\partial y^2} \end{bmatrix} \equiv \sigma^2 \begin{bmatrix} L_{xx} & L_{xy} \\ L_{yx} & L_{yy} \end{bmatrix} , \tag{3.6}$$

where $L_{xy} = L_{yx}$.

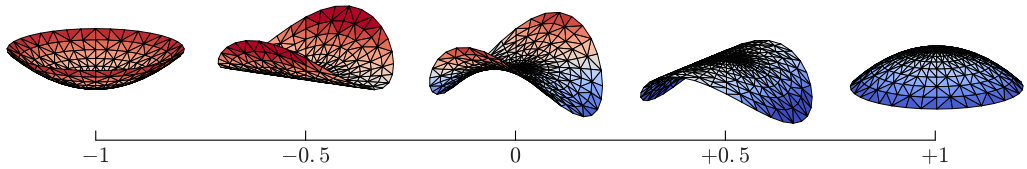


Figure 3.2 Types of surfaces corresponding to different values of the shape index. The scale at the bottom shows the shape index values. $s = -1$ corresponds to dark blobs, $s = -0.5$ to valleys, $s = 0$ to saddle points, $s = 0.5$ to ridges, and $s = 1$ to bright blobs.

Gradient orientation and magnitude

It is common to consider first order differential structure for image descriptors, that is, the gradients. Often, the gradient information is summarised in *gradient orientation* and *gradient magnitude* histograms.

The gradient orientation is simply the angle of the gradient vector,

$$\theta(x, y; \sigma) = \arctan\left(\frac{g_y}{g_x}\right) = \arctan\left(\frac{L_y}{L_x}\right), \tag{3.7}$$

and the magnitude is the length of the gradient vector,

$$M(x, y; \sigma) = \|\mathbf{g}\| = \sqrt{g_x^2 + g_y^2} = \sigma\sqrt{L_x^2 + L_y^2}, \tag{3.8}$$

where we have accounted for scale normalisation.

Shape index and curvedness

Koenderink and van Doorn (1992) defined the shape index, s, and curvedness, c, as

$$s = \frac{2}{\pi}\arctan\left(\frac{\kappa_2 + \kappa_1}{\kappa_2 - \kappa_1}\right), \quad (\kappa_1 \geq \kappa_2) \tag{3.9}$$

$$c = \sqrt{\frac{\kappa_1^2 + \kappa_2^2}{2}}, \tag{3.10}$$

where κ_1 and κ_2 are the principal curvatures at a given point on a surface. For a surface $I : \mathbb{R}^2 \to \mathbb{R}$, for instance an image, these can be found as the eigenvalues of the Hessian matrix, Eq. (3.6).

The curvedness is a positive number describing the amount of curvature, that is, how pronounced the local shape is. The shape index, on the other hand, captures the local shape as a number in the range $[-1, +1]$ and is scale invariant in the sense that scaling a structure leaves its shape index unchanged. The range encodes the local shape from spherical cups ($s = -1$) over valley-, saddle point- ($s = 0$), and ridge-like structures to spherical caps ($s = +1$), as seen in Fig. 3.2.

This encoding is particularly useful for capturing the morphology of a galaxy, as one may expect bright stars to correspond to bright blobs, $s = +1$, spiral patterns to be ridge-like structures, $s \approx +0.5$, and regions with dust to have $s < 0$, since dust will obscure light and thus appear as dark regions. Thus, the hypothesis is that a histogram of the shape index values for galaxy will correlate with the amount of dust and structure in the galaxy, in turn correlating with evolutionary parameters, such as the SFR.

We want to rewrite the shape index and curvedness in terms of image derivatives instead of eigenvalues. The eigenvalues of the Hessian matrix of the image can be written

as

$$\kappa_{1,2} = \frac{1}{2}\operatorname{Tr}\mathbf{H} \pm \frac{1}{2}\sqrt{(\operatorname{Tr}\mathbf{H})^2 - 4\det\mathbf{H}}\,, \tag{3.11}$$

meaning that

$$\kappa_2 + \kappa_1 = \operatorname{Tr}\mathbf{H} = \sigma^2(L_{xx} + L_{yy})\,, \tag{3.12}$$

$$\kappa_2 - \kappa_1 = -\sqrt{(\operatorname{Tr}\mathbf{H})^2 - 4\det\mathbf{H}} \tag{3.13}$$

$$= -\sigma^2\sqrt{(L_{xx} + L_{yy})^2 - 4(L_{xx}L_{yy} - L_{xy}^2)} \tag{3.14}$$

$$= -\sigma^2\sqrt{(L_{xx} - L_{yy})^2 + 4L_{xy}^2}\,. \tag{3.15}$$

The shape index can thus be written as

$$S(x, y; \sigma) = \frac{2}{\pi}\arctan\left(\frac{-L_{xx} - L_{yy}}{\sqrt{(L_{xx} - L_{yy})^2 + 4L_{xy}^2}}\right). \tag{3.16}$$

We can write the curvedness in terms of image derivatives in the same way. Since

$$\kappa_{1,2}^2 = \frac{1}{2}(\operatorname{Tr}\mathbf{H})^2 - \det\mathbf{H} \pm \frac{1}{2}\operatorname{Tr}\mathbf{H}\sqrt{(\operatorname{Tr}\mathbf{H})^2 - 4\det\mathbf{H}}\,, \tag{3.17}$$

the curvedness can be written as

$$C(x, y; \sigma) = \sqrt{\frac{\kappa_1^2 + \kappa_2^2}{2}} = \sqrt{\frac{(\operatorname{Tr}\mathbf{H})^2 - 2\det\mathbf{H}}{2}} = \frac{\sigma^2}{\sqrt{2}}\sqrt{L_{xx}^2 + 2L_{xy}^2 + L_{yy}^2}\,. \tag{3.18}$$

From texture descriptors to features

As we are targeting the application of estimating a galaxy's total star formation rate, the per-pixel values of the texture descriptors are not necessarily informative. Instead, we would like to summarise the entire image, in our case of a galaxy, in the form of a distribution of values. We therefore construct histograms of the computed texture values.

Using a smooth histogram, as introduced by Koenderink and Doorn (1999), provides a more robust estimate than a standard histogram, since it avoids artefacts caused by the hard binning, in particular for values close to the bin edges. They defined the smooth histogram as

$$H(i; \mathbf{r}_0, \sigma, \beta, \alpha) = \frac{1}{2\pi\alpha^2}\int A(\mathbf{r}; \mathbf{r}_0, \alpha)B(I(\mathbf{r}; \sigma); i, \beta)\,\mathrm{d}\mathbf{r}\,, \tag{3.19}$$

where A is an aperture, defined to be a Gaussian function centred on $\mathbf{r}_0 \equiv (x_0, y_0)^\mathsf{T}$ having standard deviation α, which focuses the function on different parts of the image. The bin distribution function, B, distributes the (smoothed) image values, $I(\mathbf{r}; \sigma)$, into the histogram bins centred on bin i with a 'smoothing' of width β. Koenderink and Doorn (1999) define this to be a Gaussian function,

$$B(I; i, \beta) = \exp\left(-\frac{(I(\mathbf{r}; \sigma) - i)^2}{2\beta^2}\right). \tag{3.20}$$

The integral runs over the entire image, leaving it to the aperture function to define the local region to focus on. We therefore operate on three distinct scales: that of the aperture,

$A(\mathbf{r}; \mathbf{r}_0, \alpha)$, that of the scale-space representation of the image, $I(\mathbf{r}; \sigma)$, and that of the bin distribution function, $B(I; i, \beta)$.

In our application, we extended this formulation to also contain a weight factor F:

$$H(i; \mathbf{r}_0, \sigma, \beta, \alpha) = \int F(\mathbf{r}) A(\mathbf{r}; \mathbf{r}_0, \alpha) B(I(\mathbf{r}; \sigma); i, \beta) \, \mathrm{d}\mathbf{r} \,, \tag{3.21}$$

where A is again the aperture, and B is the binning function centred on bin i. As aperture A we used a segmentation of the galaxy, separating it from the background, computed by SExtractor (Bertin and Arnouts, 1996). For binning the shape index values, we chose B to be a Gaussian distribution

$$B(S_i; \mathbf{r}, \beta) = \exp\left(-\frac{(S(\mathbf{r}; \sigma) - S_i)^2}{2\beta^2}\right) , \tag{3.22}$$

where S_i is the shape index bin, and used the curvedness, C, as the weight factor F. This weight factor will amplify the significance of the shape index in regions of the image where the local shape is particularly pronounced, thus having a large curvedness.

A Gaussian distribution is not a good choice for binning the gradient orientations, since is does not account for the fact that the angle is periodic. Instead, we propose to use a von Mises distribution, the circular analogue of the Gaussian distribution,

$$B(\theta_i; \mathbf{r}, \beta) = \exp\left(\frac{\cos(\theta(\mathbf{r}; \sigma) - \theta_i - \theta_0)}{\beta}\right) , \tag{3.23}$$

where θ_i is the gradient orientation bin and θ_0 is a fiducial orientation. The weight factor, F, was chosen to be the gradient magnitude, thus downweighing regions of the image that are close to being flat.

3.2 Feature selection

All features are equal, but some features are more equal than others, could have been a quote from a 1940s novella. It is not. But every now and then, one has to make a hard decision of which features are worth more than the rest.

In the machine learning literature, a feature is simply the vector components of a datum, such as a position or velocity vector. When estimating, say, redshift based on magnitudes or colours of a galaxy, the magnitudes or colours are features, and together form a magnitude or colour vector for that galaxy. One could also construct a colour-magnitude vector by combining all colours and magnitudes in a single vector, or, for that matter, combine all measured quantities of a galaxy into a single vector describing that galaxy. These components would all be thought of as features, and for large surveys they may add up to hundreds of features.

Data sets are getting larger and larger – not just in terms of samples, but also in terms of features. While each of these features may be informative, high dimensional spaces are problematic to work in for many methods, since data become sparse. Thus, it can be of interest to reduce the set of features to the most informative ones. Other reasons for removing less informative features may be to decrease storage requirements, or to speed up evaluation of certain algorithms. From a physical point of view, knowing the most informative features may provide insight into the problem at hand or about the process that generated the data.

One way to reduce the number of features is to use *feature extraction* methods, which try to form new, more informative features from combinations of the original ones. Principal component analysis (PCA) is a well-known feature extraction and dimensionality reduction method. One may also attempt to simply find the most informative combination of original features, a task known as *feature selection*. Guyon and Elisseeff (2003), and more

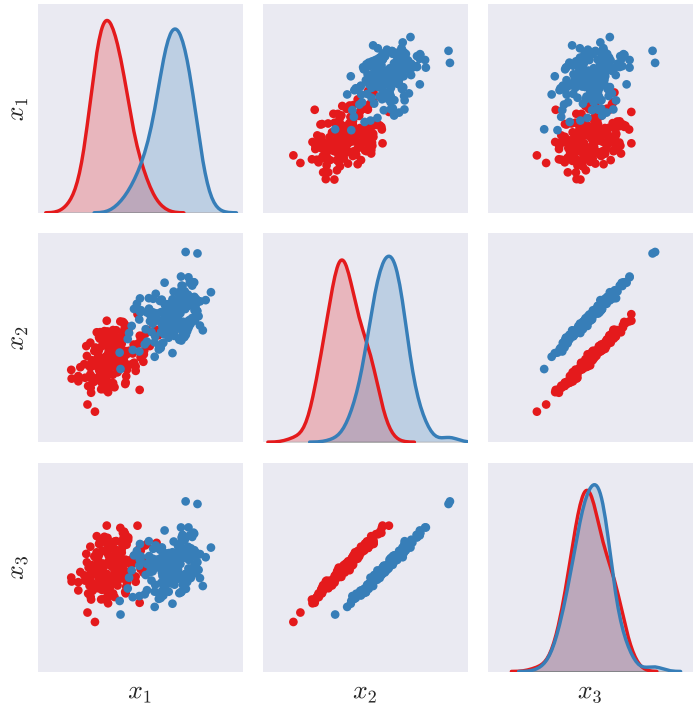


Figure 3.3 Illustration of a two-class classification problem with two 3-dimensional Gaussian distributions. Ranking the features individually by their ability to separate the two classes will show that x_1 is the most informative, while x_3 is completely useless. Thus, one could be tempted to remove x_3, as x_1, perhaps in combination with x_2, intuitively should work much better. However, combining x_2 with the presumably useless feature x_3 turns out to provide a representation where the two classes are completely separable. The mistake was of course to only judge the features by themselves, ignoring correlations with others. One should never discard a feature without a fair trial.

recently Li et al. (2016), have reviewed a number of feature selection strategies, a few of which are summarised here.

Consider a data set $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\} \in \mathbb{R}^D \times \mathbb{R}$ consisting of D-dimensional inputs $\mathbf{x}_i = (x_1, \ldots, x_D)^T$ and associated output values y_i. In the most simple feature selection setting, one may try to rank each feature x_j individually in terms of information content. While features found from such methods may, by themselves, be most discriminative, a combination of these features may not be the most discriminative combination. As illustrated in Fig. 3.3, a combination with a feature that by itself is completely useless may add significant discriminative power in combination with others, even if these are not the most expressive either.

A number of methods to alleviate the problems of single feature ranking have been proposed. According to Guyon and Elisseeff (2003), these can broadly be categorised as either *filters*, *embedded methods*, or *wrappers*.

Filters Usually applied as a preprocessing step, filters attempt to remove the least informative features before other strategies are employed. This can be necessary in situations where the number of features extends to many thousands or more. Typical filter methods first rank features (either individually or in combinations) before removing the least informative. Filters are based on various information measures, such as correlations or mu-

tual information, and are thus independent of any particular machine learning algorithm (Li et al., 2016).

Embedded methods Some machine learning algorithms have feature selection as part of their training procedure. These are called embedded methods. Such methods may be beneficial, as one gets feature rankings for free. Examples of such methods are decision trees such as CART and random forests. Kernel methods using certain kernels containing feature scaling factors, tuned as hyperparameters, may also provide feature rankings. The radial basis function kernel is an example of such a kernel, where the length scale may be interpreted as (inverse) feature importance, since long length scales mean less expressive and thus less important features (Bishop, 2009).

Wrappers Why define your own information content, when you can piggyback on machine learning algorithms? This is exactly what wrappers do; they just assess the predictive power of the algorithms when exposed to subsets of the features. This allows for particularly flexible feature selection strategies as the machine learning algorithms are defining the information content. A major problem, however, is how one chooses the feature subsets to test. Ideally, one would test all combinations, but the number of combinations increases exponentially with the number of features, quickly making such a brute force approach infeasible. Thus one needs to devise strategies for selecting subsets to test.

The two most common strategies are *forward feature selection* and *backwards feature elimination*. In forward feature selection, one iteratively tries every single feature for the prediction task. The feature yielding the best performance is kept, and one now combines it with every remaining feature in turn, assessing the performance of each pair. The best pair is kept, and the procedure is repeated. Backwards feature elimination starts from all D features, removes each feature in turn and keeps the $D-1$ features yielding the best performance. The process is repeated, each time resulting in the removal of the least informative feature.

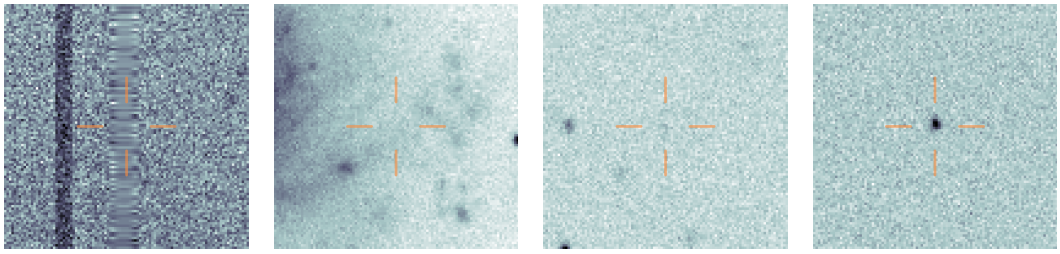
Stensbo-Smidt et al. (2017), chapter 8, investigates the hypothesis that the standard set of features for a number of estimation tasks in astrophysics, such as redshift and specific SFR estimation, may not be optimal. These features are often selected based on a combination of experience and physical knowledge, but many more features are readily available. We thus look for a combination of features from a much larger feature set yielding better performance for these two estimation tasks.

We choose forward feature selection as our strategy and combine it with k nearest neighbours (k-NN) regression to perform the estimations. k-NN is an algorithm that can be efficiently implemented on graphical processing units (GPUs), making forward feature selection a very feasible approach and also gives us a ranking of each individual feature. With such rankings one may try to interpret the importance of the features in terms of the task at hand, which would have been much more difficult with feature extraction methods, such as PCA.

3.3 Detecting distant quasars

As mentioned in section 2.3, quasars are incredibly powerful and bright objects. They are usually also very distant, meaning that despite emitting a lot of light, they appear small and faint on the sky. Detecting them is a major issue.

We here consider the problem of detecting a quasar at $z > 8$. When quasar candidates are detected in images by automatic software, they are evaluated by algorithms estimating the probability that these candidates are 'interesting', that is, whether they are indeed quasars and at the same time very distant. These algorithms do, however, occasionally get confused by nearby phenomena in the images. Examples of some of the most common problematic cases are shown in Fig. 3.4. All images are 100×100 pixel cut-outs of the larger



(a) CCD flaws. **(b)** Coinciding galaxy. **(c)** Nearby sources. **(d)** Visible source.

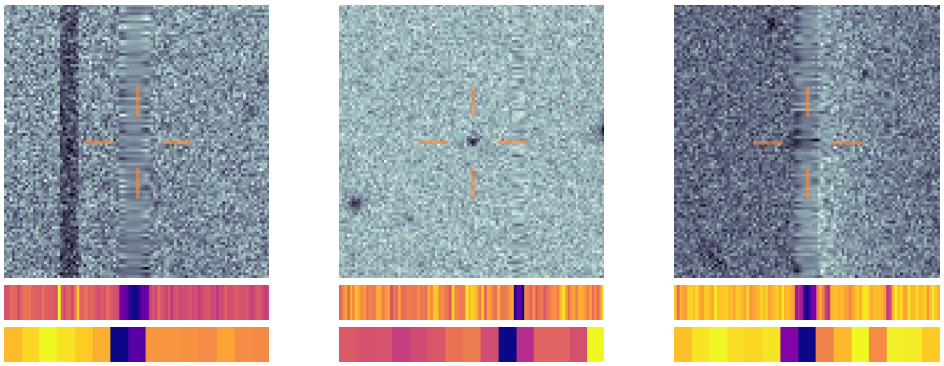
Figure 3.4 Examples of false positive quasar candidates in inverted colours obtained from SDSS. In all images the proposed candidate is exactly in the middle of the image. a shows two CCD flaws originating from a saturated nearby star, b shows a candidate which coincides with a galaxy, c shows nearby (faint) sources, and d shows a visible source even though it should not be.

fields obtained by SDSS. The cut-outs are constructed such that the quasar candidates are always centred in the cut-outs. Our task is to quality check each cut-out to make sure that there are no phenomena, physical or artificial, in the images, which would render them 'uninteresting'.

Figure 3.4a shows a candidate coinciding with a CCD flaw originating from a nearby bright star, which has saturated the CCD. Such a flaw can confuse the algorithms, even when it is just close to the candidate and not necessarily overlapping it. In such cases we do not trust the candidate, and it should be discarded. Figure 3.4b shows a candidate coinciding with a galaxy. It is highly likely that the algorithms have detected something within the galaxy rather than behind it, so the candidate should be discarded. Figure 3.4c shows sources close to the candidate source, with some of the nearby sources being very faint, for instance the source in the top right corner. With such nearby sources, the detection algorithms can get confused, or the candidate may in fact be an asteroid or comet that has moved in the time between the optical and infrared images were obtained. Thus the candidate cannot be trusted and should be discarded. Finally, Fig. 3.4d shows a candidate that is clearly visible. The images are from SDSS, an optical survey, but since we are targeting quasars at $z > 8$, the candidates will be redshifted out of the optical spectrum. Instead, detection algorithms are scanning infrared surveys for sources that are visible here, but missing in optical surveys. Thus, even if this is a quasar, it is too close to be interesting for our purpose, so the image should be discarded.

These examples, and hundreds more, were all manually inspected and discarded as part of the work done by Mortlock et al. (2011). It is difficult to automatically detect and remove these false positive candidates, since the selection criteria can be difficult to formalise and often rely on expert gut feelings. Consider for instance Fig. 3.4c. Are the nearby sources too close to the centre for us to trust the candidate? What if they were a bit brighter or a bit fainter? There also appears to be a very faint source right at the centre – is this a source or simply noise? If it is a source, is it likely to be at too low a redshift to be interesting? Such questions are difficult to create strict rules for, but we may be able to teach a computer to do this kind of assessment for us. Thus, it seems natural to turn to machine learning and image analysis for assistance.

We have tested various image filtering techniques in an attempt to detect as many of these false positives as possible. The filtered images need to be reduced to a set of features and fed to a classifier, which should then learn to distinguish true positives from false positives.



(a) (b) (c)

Figure 3.5 Finite difference applied to three images affected by the CCD flaws. The first band below the source images shows the column average of the finite difference of the image. The smeared CCD flaw is easily detected, but the dark stripe in a is not. It can, however, be detected using the Hessian eigenvalues. The second band shows the first band downsampled. It is seen that the information is still there, but the feature vector size has been significantly reduced.

Detecting CCD flaws

Finite difference seems to be a particularly efficient way of detecting CCD flaws, as seen in Fig. 3.5. The CCD flaws always occur as vertical stripes of pixels that appear to have been smeared horizontally, so using finite difference, we can subtract each pixel column from its neighbour and find the average of the column:

$$\langle |\Delta I| \rangle_i = \frac{1}{N_i} \sum_{i=1}^{N_i} |I_{i,j+1} - I_{i,j}| \quad \text{for } j = 1, \ldots, N_j - 1 \,, \tag{3.24}$$

where I is the image having N_i pixel rows and N_j pixel columns. Subtracting two neighbouring columns corresponds to calculating the gradient L_x at pixel-scale. This reduces the image to a vector of length $N_j - 1$, where each element j is the average of the differences in all rows in pixel column j. The vector can be subsampled to reduce dimensionality further.

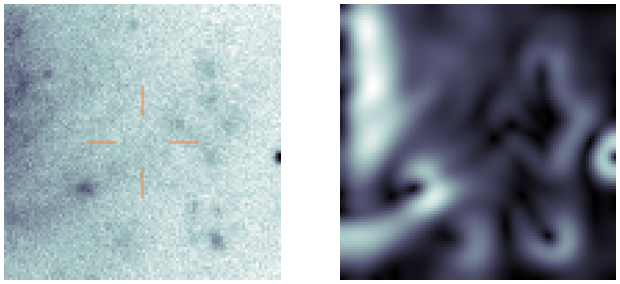
Detecting coinciding galaxies

Spiral galaxies exhibit structure, such as stars and dust, but ellipticals do not, so building a feature relying on structure will not capture all of these false positives. All galaxies will, however, cause a significant gradient in the image, so this is likely a better feature. We therefore compute the gradient magnitude, defined in Eq. (3.8), for every pixel in the cutout, see Fig. 3.6. To create a feature from this, we can either just use the average magnitude or create a histogram of all values.

Detecting (faint) sources

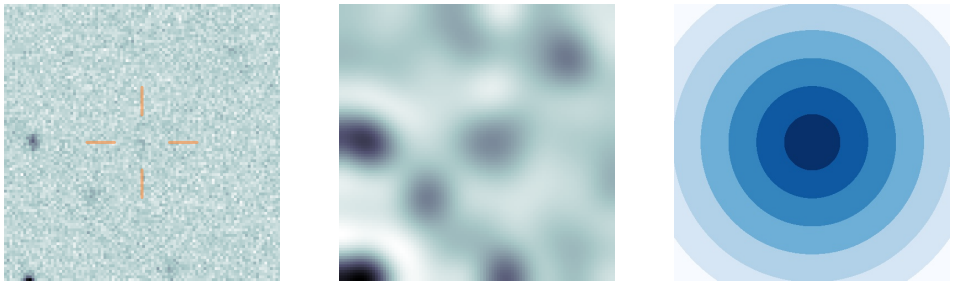
A bright source, like the one in Fig. 3.4d, is easy enough to detect; one can simply just look for large pixel values. However, many sources are very close to the background noise level, making them very difficult to detect. Nonetheless, they are important to discover before one wastes expensive and time-consuming telescope time imaging the region again.

We have found that the sum of the Hessian eigenvalues, Eq. (3.12), can trace even faint sources, see Fig. 3.7. The sum of the Hessian eigenvalues, also called the *Laplacian*, is one of the most commonly used blob detectors in computer vision.



(a) Original image. **(b)** Gradient magnitudes.

Figure 3.6 Example of gradient magnitudes computed for each pixel of an image. a shows the original image; b shows the gradient magnitudes at each pixel. For the gradient magnitudes, light colour means large gradient. It is seen that the galaxy creates large gradients in the image. The gradient magnitudes would be essentially zero for an image of pure, random noise.



(a) Original image. **(b)** Sum of Hessian eigenvalues. **(c)** Circular bands for binning.

Figure 3.7 Example of the sum of Hessian eigenvalues computed for each pixel of an image. a shows the original image, b shows the sum of Hessian eigenvalues for each pixel, and c shows the circular bands used to bin the sum of the Hessian eigenvalues for each image. The values falling in each concentric band will be assigned the same bin in the resulting histogram, and the each bin will be normalised to the amount of pixels falling in that bin. By comparing a and b, it is seen that even faint sources are quite visible after the filtering. In particular, the faint source at the top right corner and at the centre of the cut-out are clearly visible in the filtered image.

Since the closer a source, bright or faint, comes to the centre of the image, the bigger a problem it becomes, we use circular bins from the centre of the image, when constructing a histogram, see Fig. 3.7c. This makes it possible to see how far from the centre a source is, thus allowing the classifier to learn when a source comes close enough to be a problem.

As mentioned, this is still work in progress, but from the shown examples it is clear that we can detect many of the false positives created by the detection pipelines. After the filtering and extraction of features, classification of the cut-outs are done by a support vector machine (SVM). Preliminary results show an accuracy above 90% for an accept/reject classification task using roughly 1000 images. However, some of the manually assigned labels have been found to be wrong, and before we know the full extent of this mislabelling, we cannot make a proper statistical assessment.

A clear issue of the current approach is the amount of parameters that we need to tune, for instance, the scale, σ, the cut-out size, and the number of bins in the various histograms. With further work, however, we will hopefully be able to reduce these, or at least be able to provide tighter intervals on useful parameter values.

4

Summary

4.1 Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy

In Kremer et al. (2016), we provide a dissemination and introduction to the field of astroinformatics aimed at computer scientists. We motivate the need for machine learning and image analysis in astronomy, present examples of astroinformatical works, and discuss current challenges.

The paper has been accepted for publication in a special issue of IEEE Intelligent Systems. Note that the paper was limited to 5400 words and 15 references.

4.2 Shape Index Descriptors Applied to Texture-Based Galaxy Analysis

In Steenstrup Pedersen et al. (2013), we consider the problem of estimating the specific star formation rate (sSFR) of galaxies from broad-band images only. In particular, we aim to base the estimation solely on the morphology of the galaxy, thus bypassing the standard method of spectral energy distribution (SED) fitting. For describing the galaxy morphologies we propose a novel texture descriptor based on gradient orientations and the shape index.

Results show that the gradient orientations do not carry any significant information, achieving a root-mean-square error (RMSE) of $(0.81 \pm 0.02) \times 10^{-2}\,\log(\mathrm{yr}^{-1})$ compared to just predicting the average sSFR, which achieves an RMSE of $(0.88 \pm 0.02) \times 10^{-2}\,\log(\mathrm{yr}^{-1})$. The shape index does carry information, giving an RMSE of $(0.53 \pm 0.03) \times 10^{-2}\,\log(\mathrm{yr}^{-1})$, though not as much as SED-fitting, achieving an RMSE of $(0.33 \pm 0.01) \times 10^{-2}\,\log(\mathrm{yr}^{-1})$. Augmenting the output of the SED-fitting to the shape index features, however, shows an improvement in the RMSE, which now decreases to $(0.29 \pm 0.02) \times 10^{-2}\,\log(\mathrm{yr}^{-1})$. This suggests that the shape index is able to extract information not captured by the SED-fitting, though more work is needed to confirm this.

From this work, it is clear that the proposed texture descriptor using the shape index does capture information about the morphology of galaxies, relevant to the estimation of sSFRs. Further work may improve the estimations, or may show correlations with other parameters of the galaxies. Furthermore, being a novel texture descriptor, we have made an important contribution to the field of texture analysis in computer science.

4.3 Sacrificing information for the greater good: how to select photometric bands for optimal accuracy

In Stensbo-Smidt et al. (2017), we consider the problem of selecting the most informative features among all the measured magnitudes for galaxies in the SDSS database in an attempt to estimate redshift and specific star formation rate from photometry alone. We use a massively parallel implementation of the k nearest neighbours algorithm together with forward feature selection (both implemented on graphical processing units, GPUs) to do the estimations and select the best features.

Using data for 603,680 galaxies from the SDSS database, we find that for estimating the specific star formation rate, we are able to achieve a root-mean-square error (RMSE) of $(27.4 \pm 0.3) \times 10^{-2} \log(\mathrm{yr}^{-1})$ when using optimal features. This should be compared to an RMSE of $(29.6 \pm 0.2) \times 10^{-2} \log(\mathrm{yr}^{-1})$ when using the four `modelMag` colours as features, as often seen in the astronomical literature and as advocated by SDSS.

For the task of photometric redshift estimation, we achieve a normalized median absolute deviation, σ_{NMAD} of $(1.38 \pm 0.01) \times 10^{-2}$ when using optimal features. SDSS achieve a σ_{NMAD} of $(1.65 \pm 0.01) \times 10^{-2}$ for the same galaxies. We thus significantly outperform SDSS for the task of photometric redshift estimation.

The results show that there can be significant gains in accuracy for different tasks by not relying solely on the magnitudes recommended in standard literature. Whereas our method does select the `modelMag` colours as some of the most informative, adding additional colours and magnitudes clearly improved the results.

The suggested feature selection method is completely general and can thus help astronomy and astrophysics in many aspects. Furthermore, the highly optimized GPU implementation of k nearest neighbours and forward feature selection can be directly used in many areas of computer science.

4.4 Automating the quality assessment of images of distant quasar candidates

Though still work in progress, the project on quality checking images of distant quasar candidates show promising results. We consider the problem of assessing whether an image of a candidate for a distant quasar meet certain quality criteria set by astrophysicists. The problem is further complicated by the fact that these criteria can be difficult, if not impossible, to define in a strict mathematical sense. We therefore seek to extract features from the image using image analysis and use machine learning to train a program to imitate human decision making.

We are able to detect the most common causes of images having to be discarded as a result of either physical or artificial phenomena in the image. We currently lack a data set of sufficient quality to make a proper statistical assessment of our method, but preliminary results suggest that we are able to correctly accept or reject an image in more than 90% of the cases.

5

Perspectives and future work

Astroinformatics is a rapidly growing research domain, which this thesis has contributed to. New methods that can improve measurements in astronomy and astrophysics have been introduced, leading the ways for new discoveries. Contributions to computer science have been made by the means of a novel texture descriptor and adaptation of known methods to fit problems of astronomical interest. There are, of course, plenty of ways to extend and build on the knowledge we have gained.

5.1 Morphology of galaxies

Steenstrup Pedersen et al. (2013) considered the problem of describing the morphology of galaxies in terms of texture and correlate this with their specific star formation rate. We found that the shape index and curvedness provided additional information about the star formation rate not detected with standard SED fitting. This result suggests that the texture descriptors could maybe also be useful for describing other physical processes within galaxies, or perhaps be used to study the physics and structure of nebulae.

We chose to use 'handcrafted' features for describing the texture rather than using the ubiquitous convolutional neural network (CNN) approach for a number of reasons. Firstly, manually constructed features are often easier to interpret, and in our application we can directly interpret the shape index in terms of physical structure. Interpreting the representations learnt by CNNs can be difficult, as they are often complex combinations of simpler features. Secondly, training a CNN to estimate, say, specific star formation rates from images will tune it to focus on texture relevant for this particular purpose. Manually constructed features will be more general, and may therefore be used in more diverse settings, including unsupervised ones. For instance, assuming the shape index traces some fundamental structure in galaxies, one could now look for clusters in 'shape index space', which could potentially correspond to fundamentally different types of galaxies. Thus, one could construct a new, more data-driven classification scheme to replace the century old Hubble classes.

This kind of texture analysis can therefore be highly promising with regards to gaining new, physical insights. Also computer science will benefit from this approach as it, like in this case, often involves inventing new, interesting methods.

That being said, using manual features are not without complications. The shape index and curvedness needs to be calculated at different scales, in order to pick up information from different sized structures. Also, the histograms used depended on a few parameters, such as the number of bins and the width of the binning function. All these parameters, and the number and location of the scales at which to compute the shape index and curvedness, need to be determined. Currently, they are just set by hand to reasonable values, but it would of course be good to do a proper test to find optimal values.

It may be possible to learn some of the parameter values directly from data. For instance, it may be that galaxies typically show structures only on a fixed number of scales, although these scales will vary from image to image, depending on the apparent size of the galaxy. Thus, learning the amount of informative scales may be possible using the entire data set, whereas their exact locations will be image-specific.

Selecting which scales to use is a particularly interesting problem to spend more time on. Scale selection in general is an unsolved problem, as it requires us to locate the most informative scales. Searching through scale space is computationally intensive, and even the notion of 'informative' is not clear. There have been attempts to formalise the notion of an informative scale (e.g. Lindeberg, 1998; Sporring and Weickert, 1999), and investigating these would be a natural next step.

The current formulations of shape index and curvedness also only work on greyscale images. We circumvented this limitation by concatenating histograms computed for each SDSS band, but this leads to an increase in the number of features, which can decrease estimation accuracy. Additionally, we would expect the light in different bands to be highly correlated, so we may be adding somewhat redundant features. Extending the shape index and curvedness to handle multicolour images properly could give us much more expressive features.

Lastly, one could also work on the segmentation of the galaxies and how that is used in the further computations. We used a segmentation based on the output from SExtractor (Bertin and Arnouts, 1996) and computed a single histogram for the entire galaxy. One could imagine other approaches, such as computing histograms for different parts of the galaxies, or weighing different parts differently. In particular, it is difficult to define the edge of a galaxy so downweighing pixels close to the assumed edge could make sense. We already do something along these lines, using the segmentation found by SExtractor, but investigating other approaches would be natural.

5.2 Selecting informative features

Stensbo-Smidt et al. (2017) introduced a general method for selecting the most informative features for a given task. We exemplified its usefulness by estimating redshifts and specific star formation rates using features selected for the tasks, demonstrating an increase in accuracy compared to using more standard features. Since the method is completely general, it can help in other areas of astronomy and astrophysics as well.

The method we used, forward feature selection, is one of the simplest feature selection algorithms available, and choosing this particular method was a deliberate choice. We wanted to show that even a simple method can lead to improvements and interpretable results. One could, of course, try other methods that might perform even better. The tricky part is to find methods that can deal with the massive amounts of data available in astronomy. We solved this by creating a clever data structure that allowed the feature selection task to be run on GPUs, ensuring massive gains in speed.

An interesting extension to our work would be to include support for uncertainties, both on the features and on the targets. This is not easily done, but it may be possible using a Bayesian framework. One could also investigate various heuristics for downweighing the importance of uncertain inputs and for estimating the error on the predicted target.

5.3 Detecting distant quasars

Finally, we have considered the problem of automatically assessing the quality of images of quasar candidates. While the work here is still ongoing, there are some interesting directions we could take. Currently, we are targeting a simple accept/reject scheme, assessing whether a candidate should be kept or discarded. It would be interesting to instead consider a probabilistic multiclass classification scheme, assessing the probability that a can-

didate belongs to each of the many different classes of false positives. This would allow us to set a threshold for when we would like to do manual inspection, a threshold that could vary from class to class.

A significant challenge is to account for unknown classes of false positives as they are discovered. We may need to construct new features to be able to detect these, and we also need to retrain the classifier. With the amount of parameters we can tune, finding ways to limit the amount of needed retraining when adding new classes is a crucial task.

Another interesting and promising path is to consider semi-supervised learning. Semi-supervised learning is a type of supervised learning, which tries to also use unlabelled data to help inform the supervised task during training. This is an interesting approach because of the large amount of manual work required to create a large training set of images of candidates. All images must be manually labelled by an expert, which is a tedious and slow process. SDSS offers terabytes of images from large regions of the sky, so we can easily get thousands of images to train the classifier on – we just don't know whether these images would be considered good or bad, thus they are unlabelled. A semi-supervised algorithm could potentially learn the general structure of a patch of sky, and then use this information to make better use of manually inspected candidates, as demonstrated by Kingma et al. (2014), although for much simpler data sets.

Software that can automatically assess the quality of images will of course not only be of use for quasar detection. Any search for rare astrophysical phenomena will need automatic quality checks simply because of the rare nature of the phenomena; there will be so many false positives that humans cannot assess them all. Computer science will benefit through the experience gained by adapting existing methods and developing new ones to solve problems faced in this new domain.

5.4 The relevance of interdisciplinary research

Astronomy and computer science can gain a lot from the mutual collaboration of astroinformatics. Interdisciplinary research is, in my experience, highly rewarding, but at the same time difficult. Coming from different fields and backgrounds, we as scientists are raised with a specific mindset and language. This is not something one often thinks about, but it becomes very apparent when doing interdisciplinary research. There is a clash of cultures and traditions, which can lead to frustrating discussions and heated arguments, but this is a good thing; we are forced to reconsider ways of thinking that we have taken for granted, which is very educational. We also appreciate domain expertise vastly more, in particular after subtle details in, for instance, the used data that have been overlooked makes your entire experiment void.

There are indeed headaches associated with doing interdisciplinary research, but there is also a wealth of knowledge and mutual respect to be gained. I firmly believe that the scientific breakthroughs of the future will happen through the joined forces of distinct research fields.

Part II

Included papers

6

Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy

Bibliographic reference

J. Kremer, K. Stensbo-Smidt, F. Gieseke, K. Steenstrup Pedersen, and C. Igel. Big universe, big data: machine learning and image analysis for astronomy. *IEEE Intelligent Systems*. Accepted for publication, September 2016.

Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy

Jan Kremer[*], Kristoffer Stensbo-Smidt[*], Fabian Gieseke[**], Kim Steenstrup Pedersen[*], and Christian Igel[*]

[*]Department of Computer Science, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen Ø, Denmark
[**]Institute for Computing and Information Sciences, Radboud University Nijmegen, Toernooiveld 212, 6525 AJ Nijmegen, The Netherlands

Astrophysics and cosmology are rich with data. The advent of wide-area digital cameras on large aperture telescopes has led to ever more ambitious surveys of the sky. Data volumes of entire surveys a decade ago can now be acquired in a single night and real-time analysis is often desired. Thus, modern astronomy requires big data know-how, in particular it demands highly efficient machine learning and image analysis algorithms. But scalability is not the only challenge: Astronomy applications touch several current machine learning research questions, such as learning from biased data and dealing with label and measurement noise. We argue that this makes astronomy a great domain for computer science research, as it pushes the boundaries of data analysis. In the following, we will present this exciting application area for data scientists. We will focus on exemplary results, discuss main challenges, and highlight some recent methodological advancements in machine learning and image analysis triggered by astronomical applications.

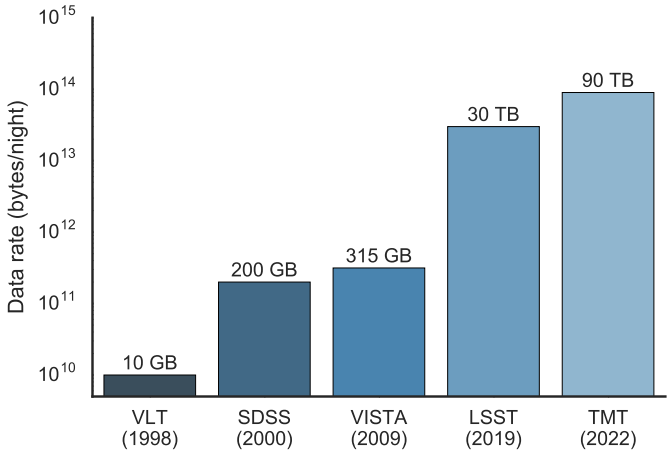


Figure 1: Increasing data volumes of existing and upcoming telescopes: Very Large Telescope (VLT), Sloan Digital Sky Survey (SDSS), Visible and Infrared Telescope for Astronomy (VISTA), Large Synoptic Survey Telescope (LSST) and Thirty Meter Telescope (TMT).

Ever-Larger Sky Surveys

One of the largest astronomical surveys to date is Sloan Digital Sky Survey (SDSS, `http://www.sdss.org`). Each night, the SDSS telescope produces 200 GB of data and now provides close to a million field images, in which more than 200 million galaxies, and even more stars, have been detected. Upcoming surveys will provide far greater data volumes.

Another promising future survey is the *Large Synoptic Survey Telescope* (LSST). It will deliver wide-field images of the sky, exposing galaxies that are too faint to be seen today. A main objective of LSST is to discover *transients*, objects that change brightness over time-scales of seconds to months. These changes are due to a plethora of reasons; some may be regarded as uninteresting while others will be extremely rare events, which cannot be missed. LSST is expected to see millions of transients per night, which need to be detected in real-time to allow for follow-up observations. With staggering 30 TB of images being produced per night, efficient and accurate detection will be a major challenge. Figure 1 shows how data rates have increased and will continue to increase as new surveys are initiated.

What do the data look like? Surveys usually make either *spectroscopic* or *photometric* observations, see Figure 2. Spectroscopy measures the photon count at thousands of wavelengths. The resulting spectrum allows for identifying chemical components of the observed object and thus enables determining many interesting properties. Photometry takes images using a CCD, typically acquired through only a handful of broad-band filters, making photometry much less informative than spectroscopy.

While spectroscopy provides measurements of high precision, it has two drawbacks: First, it is not as sensitive as photometry, meaning that distant or otherwise faint objects cannot be measured. Second, only few objects can be captured at the same time, making it more expensive than photometry, which allows for acquiring images of thousands of objects in a single image. Photometry can capture objects that may be ten times fainter than what can be measured with spectroscopy. A faint galaxy is often more distant than a bright one—not just in space, but also in time. Discovering faint objects therefore offers the potential of looking further back into the history of the Universe, over time-scales of billions of years. Thus, photometric observations are invaluable to cosmologists, as they help understanding the early Universe.

Once raw observations have been acquired, a pipeline of algorithms needs to extract information from them. Much image-based astronomy currently relies to some extent on visual inspection. A wide range of measurements are still carried out by humans, but need to be addressed by automatic image analysis in light of growing data volumes. Examples are 3D orientation and chirality of galaxies, and detection of large-scale features, such as jets and streams. Challenges in these tasks include image artifacts, spurious effects, and discerning between merging galaxy pairs and galaxies that happen to overlap along the line of sight. Current survey pipelines often have trouble correctly identifying these types of problems, which then propagate into the databases.

A particular challenge is that cosmology relies on scientific analyses of long-exposure images. As such, the interest in image analysis techniques for preprocessing and de-noising is naturally great. This is particularly important for the detection of faint objects with very low signal-to-noise ratios. Automatic object detection is vital to any survey pipeline, with reliability and completeness being essential metrics. Completeness refers to the amount of detected objects, whereas reliability measures how many of the detections are actual objects. Maximizing these metrics requires advanced image anal-

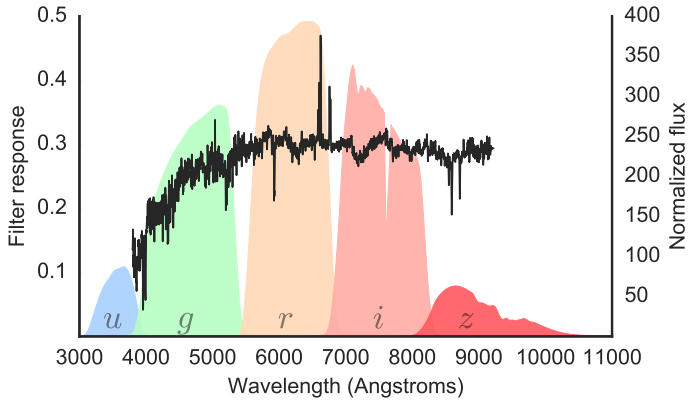


Figure 2: The spectrum of galaxy NGC 5750 (black line), as seen by SDSS, with the survey's five photometric broad-band filters u, g, r, i, and z, ranging from ultraviolet (u) to near-infrared (z). For each band the galaxy's brightness is captured in an image.

ysis and machine learning techniques. Therefore, data science for astronomy is a quickly evolving field gaining more and more interest. In the following, we will highlight some of its success stories and open problems.

Large-scale Data Analysis in Astronomy

Machine learning methods are able to uncover the relation between input data (e.g., galaxy images) and outputs (e.g., physical properties of galaxies) based on input-output samples, and they have already proved successful in various astrophysical contexts. For example, Mortlock et al.[8] use Bayesian analysis to find the most distant quasar to date. These are extremely bright objects forming at the center of large galaxies and are very rare. Bayesian comparison has helped scientists to select a few most likely objects for re-observation from thousands of candidates.

In astronomy, distances from Earth to galaxies are measured by their redshifts, but accurate estimations need expensive spectroscopy. Getting accurate redshifts from photometry alone is an essential, but unsolved task, for which machine learning methods are widely applied.[2] However, they are far from on a par with spectroscopy. Thus, better and faster algorithms are much desired.

Figure 3: An example of two morphology categories: on the left, the spiral galaxy M101; on the right, the elliptical galaxy NGC 1132 (credit: NASA, ESA, and the Hubble Heritage Team (STScI/AURA)-ESA/Hubble Collaboration).

Another application is the measurement of galaxy morphologies. Usually, one assigns a galaxy a class based on its appearance (see Figure 3), traditionally using visual inspection. Lately, this has been accelerated by the citizen science project *Galaxy Zoo*,[7] which aims at involving the public in classifying galaxies. Volunteers have contributed more than 100 million classifications, which allow astrophysicists to look for links between the galaxies' appearance (morphology) and internal and external properties. A number of discoveries have been made through the use of data from Galaxy Zoo, and the classifications have provided numerous hints to the correlations between various processes governing galaxy evolution. A galaxy's morphology is difficult to quantize in a concise manner, and automated methods are high on the wish list of astrophysicists. There exists some work on reproducing the classifications using machine learning alone,[3] but better systems will be necessary when dealing with the data products of next-generation telescopes.

A growing field in astrophysics is the search for planets outside our solar system (exoplanets). NASA's Kepler spacecraft has been searching for exoplanets since 2009. Kepler is observing light curves of stars, that is, measuring a star's brightness at regular intervals. The task is then to look for changes in the brightness indicating that a planet may have moved in front of it. If that happens with regular period, duration and decrease in brightness, the source is likely to be an exoplanet. While there is automated

software detecting such changes in brightness, the citizen science project *Planet Hunters* has shown that the software does miss some exoplanets. Also, detecting Earth-sized planets, arguably the most interesting, is notoriously difficult, as the decrease in brightness can be close to the noise level. For next-generation space telescopes, such as Transiting Exoplanet Survey Satellite (TESS), scheduled for launch in 2017, algorithms for detecting exoplanets need to be significantly improved to more reliably detect Earth-sized exoplanet candidates for follow-up observations.

There are also problems that may directly affect our lives here on Earth, such as solar eruptions that, if headed towards Earth, can be dangerous to astronauts, damage satellites, affect airplanes and, if strong enough, cause severe damage to electrical grids. A number of spacecrafts monitor the Sun in real-time. While the ultimate goal is a better understanding of the Sun, the main reason for real-time monitoring is to be able to quickly detect and respond to solar eruptions. The continuous monitoring is done by automated software, but not all events are detected.[13] Solar eruptions are known to be associated with sunspots, but the connection is not understood well enough that scientists can predict the onset or magnitude of an eruption. There may be a correlation with the complexity of the sunspots, and understanding this, as well as how the complexity develops over time, is crucial for future warning systems. While scientists are working towards a solution, for example through the citizen science project *Sunspotter* (`https://www.sunspotter.org/`), no automated method has yet been able to reliably and quantitatively measure the complexity.

This glimpse of success stories and open problems of big data analysis in astronomy is by no means exhaustive. An overview of machine learning in astronomy can be found in the survey by Ball and Brunner.[1]

Astronomy Driving Data Science

In the following, we present three examples from our own work showing how astronomical data analysis can trigger methodological advancements in machine learning and image analysis.

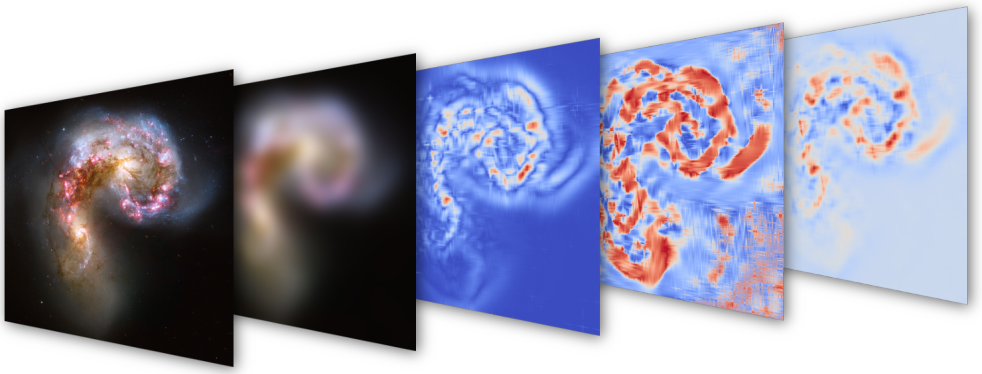


Figure 4: From left to right: The original image of a galaxy merger, the scale-space representation of the galaxies, the curvedness (a measure of how pronounced the local structure is), the shape index, and finally the shape index weighted by the curvedness. The shape index is defined as $S(x, y; \sigma) = \frac{2}{\pi} \tan^{-1}\left(\frac{-L_{xx}-L_{yy}}{\sqrt{2L_{xy}^2+(L_{xx}-L_{yy})^2}}\right)$, where $L_{x^n y^m}(x, y; \sigma) = \left(I * \frac{\partial^{(n+m)} G}{\partial x^n \partial y^m}\right)(x, y; \sigma)$ is the scale space representation of the image I, G is a Gaussian filter and σ is the scale. The curvedness is defined as $C(x, y; \sigma) = \frac{1}{2}\sigma^2\sqrt{L_{xx}^2 + 2L_{xy}^2 + L_{yy}^2}$. The image shows the Antennae galaxies as seen by the Hubble Space Telescope (credit: NASA, ESA, and the Hubble Heritage Team (STScI/AURA)-ESA/Hubble Collaboration).

Describing the Shape of a Galaxy

Image analysis does not only allow for automatic classification, but can also inspire new ways to look at morphology.[9;11] For instance, we examined how well one of the most fundamental measures of galaxy evolution, the star-formation rate, could be predicted from the *shape index*. The shape index measures the local structure around a pixel going from dark blobs over valley-, saddle point- and ridge-like structures to white blobs. It can thus be used as a measure of the local morphology on a per-pixel scale, see Figure 4. The study showed that the shape index does indeed capture some fundamental information about galaxies, which is missed by traditional methods. Adding shape index features resulted in a 12% decrease in root-mean-square error (RMSE).

Dealing with Sample Selection Bias

In supervised machine learning, models are constructed based on labeled examples, that is, observations (e.g., images, photometric features) together with their outputs (also referred to as labels, e.g., the corresponding redshift or galaxy type). Most machine learning algorithms are built on the assumption that training and future test data follow the same distribution. This allows for generalization, enabling the model built from labeled examples in the training set to accurately predict target variables in an unlabeled test set. In real-life applications this assumption is often violated—we refer to this as sample selection bias. Certain examples are more likely to be labeled than others due to factors like availability or acquisition cost regardless of their representation in the population. Sample selection bias can be very pronounced in astronomical data,[12] and machine learning methods have to address this bias to achieve good generalization. Often only training data sets from old surveys are initially available, while upcoming missions will probe never-before-seen regions in the astrophysical parameter space.

To correct the sample selection bias, we can resort to a technique called *importance-weighting*. The idea is to give more weight to examples in the training sample which lie in regions of the feature space that are underrepresented in the test sample and, likewise, give less weights to examples whose location in the feature space is overrepresented in the test set. If these weights are estimated correctly, the model we learn from the training data is an unbiased estimate of the model we would learn from a sample that follows the population's distribution. The challenge lies in estimating these weights reliably and efficiently. Given a sufficiently large sample, a simple strategy can be followed: Using a nearest neighbor-based approach, we can count the number of test examples that fall within a hypersphere whose radius is defined by the distance to the Kth neighbor of a training example. The weight is then the ratio of the number of these test examples over K. This flexibly handles regions which are sparse in the training sample. In the case of redshift estimation, we could alleviate a selection bias by utilizing a large sample of photometric observations to determine the weights for the spectroscopically confirmed training set.[6]

To measure how well we approximated the true weight we used the squared difference between true and estimated weight, that is,

$$L(\beta, \widehat{\beta}) = \sum_{x \in \mathcal{S}_{\text{train}}} \big(\beta(x) - \widehat{\beta}(x)\big)^2 p_{\text{train}}(x)\,\mathrm{d}x \ ,$$

where $\mathcal{S}_{\text{train}}$ is the training sample, β and $\widehat{\beta}$ are true and estimated weight, respectively, and p_{train} is the training density. The nearest neighbor estimator achieved similar or lower error compared to other methods. At the same time the estimator's running time is three orders of magnitude lower than the best competitor for lower sample sizes. Furthermore, it is able to scale up to millions of examples (code is available at `https://github.com/kremerj/nnratio`).

Scaling-up Nearest Neighbor Search

Nearest neighbor methods are not only useful for addressing sample selection bias, they also provide excellent prediction results in astrophysics and cosmology. For example, they are used to generate candidates for quasars at high redshift.[10] Such methods work particularly well when the number of training examples is high and the input space is low-dimensional. This makes them a good choice for analyzing large sky surveys where objects are described by photometric features (e.g., the five broad-band filters shown in Figure 2). However, searching for nearest neighbors becomes a computational bottleneck in such big data settings.

To compute nearest neighbors for a given query, search structures such as k-d trees are an established way to accelerate the search. If input space dimensionality is moderate (say, below 30), runtime can often be reduced by several orders of magnitude. While approximate schemes are valuable alternatives, one is usually interested in exact nearest neighbor search for astronomical data. In this context, massively-parallel devices, such as graphics processing units (GPUs), show great promise. Unfortunately, nearest neighbor search based on spatial data structures cannot be parallelized in an obvious way for these devices. To this end, we developed a new tree structure that is more amenable to massively-parallel traversals via GPUs, see Figure 5.[4] The framework can achieve a significant runtime reduction at a much lower cost compared to traditional parallel architectures (code available on `http://bufferkdtree.readthedocs.io`). We expect such scalable approaches to be crucial for upcoming data-intensive analyses in astronomy.

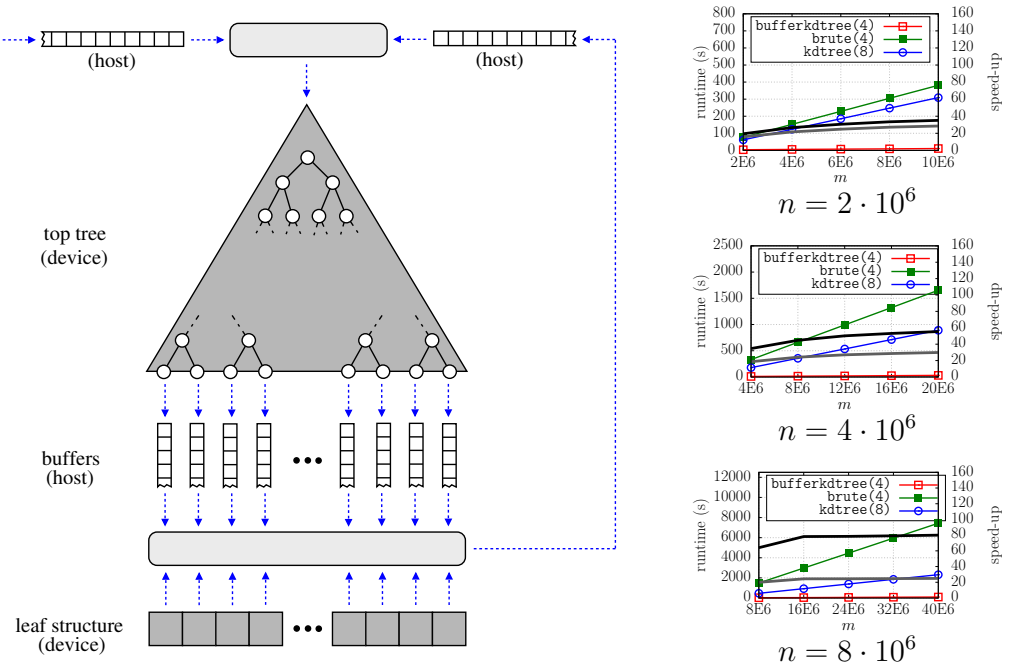


Figure 5: Left: The *buffer k-d tree* structure depicts an extension of classical k-d trees and can be used to efficiently process huge amounts of nearest neighbor queries using GPUs.[4] Right: Runtime comparison given a large-scale astronomical data set with n training and m test examples. The speed-up of the buffer k-d tree approach using four GPUs over two competitors (brute-force on GPUs and a multi-core k-d tree based traversal using 4 cores/8 hardware threads) is shown as solid black lines.[5]

Physical Models vs. Machine Learning Models

A big concern data scientists meet when bringing forward data-driven machine learning models in astrophysics and cosmology is lack of interpretability. There are two different approaches to predictive modeling in astronomy: physical modeling and data-driven modeling. Building physical models, which can incorporate all necessary astrophysical background knowledge, is the traditional approach. These models can be used for prediction, for example, by running Monte Carlo simulations. Ideally, this approach ensures that the predictions are physically plausible. In contrast, extrapolations by purely data-driven machine learning models may violate physical laws. Another decisive feature of physical models is that they allow for understanding and explaining observations. This interpretability of predictions is typically not provided when using a machine learning approach.

Physical models have the drawbacks that they are difficult to construct and that inference may take a long time (e.g., in the case of Monte Carlo simulations). Most importantly, the quality of the predictions depends on the quality of the physical model, which is typically limited by necessary simplifications and incomplete scientific knowledge. In our experience, data-driven models typically outperform physical models in terms of prediction accuracy. For example, a simple k nearest neighbors model can reduce the RMSE by 22% when estimating star formation rates.[14;15] Thus, we strongly advocate data-driven models when accurate predictions are the main objective. And this is indeed often the case, for example, if we want to estimate properties of objects in the sky for quickly identifying observations worth a follow-up investigation or for conducting large-scale statistical analyses.

Generic machine learning methods are not meant to replace physical modeling, because they typically do not provide scientific insights beyond the predicted values. Still, we argue that if prediction accuracy is what matters, one should favor the more accurate model, whether it is interpretable or not. While the black-and-white portrayal of the two approaches may help to illustrate common misunderstandings between data scientists and physicists, it is of course shortsighted. Physical and machine learning modeling are not mutually exclusive: Physical models can inform machine learning algorithms, and machine learning can support physical modeling. A simple example of the latter is using machine learning to estimate error residuals of a physical model.[9]

Dealing with uncertainties is a major issue in astronomical data anal-

ysis. Data scientists are asked to provide error bars for their predictions and have to think about how to deal with input noise. In astronomy, both input and output data have (non-Gaussian) errors attached to them. Often these measurement errors have been quantified (e.g., by incorporating weather conditions during observation), and it is desirable to consider these errors in the prediction. Bayesian modeling and Monte Carlo methods simulating physical models offer solutions, however, often they do not scale for big data. Alternatively, one can modify machine learning methods to process error bars, as attempted for nearest neighbor regression by modifying the distance function.[10]

Getting Started on Astronomy and Big Data

Most astronomical surveys make their entire data collection, including derived parameters, available online in the form of large data bases. These provide entry points for the computer scientist wanting to get engaged in astronomical research. In the following, we highlight three resources for getting started on tackling some of the open problems mentionend earlier.

The Galaxy Zoo website (`https://www.galaxyzoo.org`) provides data with classifications of about one million galaxies. It is an excellent resource for developing and testing image analysis and computer vision algorithms for automatic classifications of galaxies.

Much of the Kepler data for exoplanet discovery is publicly available through Mikulski Archive for Space Telescopes (`http://archive.stsci.edu/kepler`). These include light curves for confirmed exoplanets and false positives, making it a valuable dataset for testing detection algorithms.

Having being monitored continuously for years, there is an incredible amount of imaging data for the Sun, from archival data to near real-time images. One place to find such is Debrecen Sunspot Data archive (`http://fenyi.solarobs.unideb.hu/ESA/HMIDD.html`). These images allow for the development and testing of new complexity measures for image data or solar eruption warning systems.

A Peek Into the Future

Within the next few years, image analysis and machine learning systems that can process terabytes of data in near real-time with high accuracy will be essential.

There are great opportunities for making novel discoveries, even in databases that have been available for decades. The volunteers of Galaxy Zoo have demonstrated this multiple times by discovering structures in SDSS images that have later been confirmed to be new types of objects. These volunteers are not trained scientists, yet they make new scientific discoveries.

Even today, only a fraction of the images of SDSS have been inspected by humans. Without doubt, the data still hold many surprises, and upcoming surveys, such as LSST, are bound to image previously unknown objects. It will not be possible to manually inspect all images produced by these surveys, making advanced image analysis and machine learning algorithms of vital importance.

One may use such systems to answer questions like how many types of galaxies there are, what distinguishes the different classes, whether the current classification scheme is good enough, and whether there are important sub-classes or undiscovered classes. These questions require data science knowledge rather than astrophysical knowledge, yet the discoveries will still help astrophysics tremendously.

In this new data-rich era, astronomy and computer science can benefit greatly from each other. There are new problems to be tackled, novel discoveries to be made, and above all, new knowledge to be gained in both fields.

References

[1] N. M. Ball and R. J. Brunner. Data mining and machine learning in astronomy. *International Journal of Modern Physics D*, 19(07):1049–1106, 2010.

[2] A. A. Collister and O. Lahav. ANNz: estimating photometric redshifts using artificial neural networks. *PASP*, 116(818):345, 2004.

[3] S. Dieleman et al. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *MNRAS*, 450:1441–1459, 2015.

[4] F. Gieseke et al. Buffer k-d trees: Processing massive nearest neighbor queries on GPUs. *JMLR W&CP*, 32(1):172–180, 2014.

[5] F. Gieseke et al. Bigger Buffer k-d Trees on Multi-Many-Core Systems. In *Workshop on Big Data & Deep Learning in HPC*, 2016, in print.

[6] J. Kremer et al. Nearest neighbor density ratio estimation for large-scale applications in astronomy. *Astronomy and Computing*, 12:67–72, 2015.

[7] C. J. Lintott et al. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *MNRAS*, 389: 1179–1189, 2008.

[8] D. J. Mortlock et al. A luminous quasar at a redshift of z = 7.085. *Nature*, 474(7353):616–619, 2011.

[9] K. S. Pedersen et al. Shape Index Descriptors Applied to Texture-Based Galaxy Analysis. In *ICCV*, pages 2440–2447, 2013.

[10] K. Polsterer et al. Finding new high-redshift quasars by asking the neighbours. *MNRAS*, 428(1):226–235, 2013.

[11] K. Polsterer et al. Automatic classification of galaxies via machine learning techniques: Parallelized Rotation/Flipping INvariant Kohonen Maps (PINK). In *ADASS XXVI*, pages 81–86, 2015.

[12] J. W. Richards et al. Active learning to overcome sample selection bias: Application to photometric variable star classification. *ApJ*, 744 (2), 2012.

[13] E. Robbrecht and D. Berghmans. Automated recognition of coronal mass ejections (CMEs) in near-real-time data. *Astronomy & Astrophysics*, 425:1097–1106, 2004.

[14] K. Stensbo-Smidt et al. Nearest Neighbour Regression Outperforms Model-based Prediction of Specific Star Formation Rate. In *IEEE Big Data*, pages 141–144, 2013.

[15] K. Stensbo-Smidt et al. Simple, fast and accurate photometric estimation of specific star formation rate. *MNRAS*, 2016. Accepted subject to "moderate revisions".

Jan Kremer is a PhD candidate at DIKU, the Department of Computer Science, University of Copenhagen. His research interests include machine learning and computer vision. He has an MSc in computer science from the Technical University of Munich. Contact him at jan.kremer@di.ku.dk.

Kristoffer Stensbo-Smidt is a PhD candidate at DIKU. His research interests include statistical data analysis and astrophysics. He has an MSc in physics and astronomy from the University of Copenhagen. Contact him at k.stensbo@di.ku.dk.

Fabian Gieseke is a postdoctoral researcher at the Institute for Computing and Information Sciences, Radboud University Nijmegen. He received his PhD degree in computer science from the University of Oldenburg. His main research interests lie in the field of big data analytics. Contact him at fgieseke@cs.ru.nl.

Kim Steenstrup Pedersen is an associate professor at DIKU. He received his PhD degree from the University of Copenhagen. His research interests include computer vision and image analysis. Contact him at kimstp@di.ku.dk.

Christian Igel is a professor at DIKU. He received his Doctoral degree from Bielefeld University, and his Habilitation degree from Ruhr-University Bochum. His main research area is machine learning. Contact him at igel@diku.dk.

7

Shape Index Descriptors Applied to Texture-Based Galaxy Analysis

Bibliographic reference

K. Steenstrup Pedersen, K. Stensbo-Smidt, A. Zirm, and C. Igel. Shape index descriptors applied to texture-based galaxy analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2440–2447, 2013.

Shape Index Descriptors Applied to Texture-Based Galaxy Analysis

Kim Steenstrup Pedersen†, Kristoffer Stensbo-Smidt†, Andrew Zirm‡, Christian Igel†
† Department of Computer Science, and ‡ Dark Cosmology Centre, Niels Bohr Institute
University of Copenhagen, Denmark
http://image.diku.dk/MLLab/SkyML.php

Abstract

A texture descriptor based on the shape index and the accompanying curvedness measure is proposed, and it is evaluated for the automated analysis of astronomical image data. A representative sample of images of low-redshift galaxies from the Sloan Digital Sky Survey (SDSS) serves as a testbed. The goal of applying texture descriptors to these data is to extract novel information about galaxies; information which is often lost in more traditional analysis. In this study, we build a regression model for predicting a spectroscopic quantity, the specific star-formation rate (sSFR). As texture features we consider multi-scale gradient orientation histograms as well as multi-scale shape index histograms, which lead to a new descriptor. Our results show that we can successfully predict spectroscopic quantities from the texture in optical multi-band images. We successfully recover the observed bi-modal distribution of galaxies into quiescent and star-forming. The state-of-the-art for predicting the sSFR is a color-based physical model. We significantly improve its accuracy by augmenting the model with texture information. This study is the first step towards enabling the quantification of physical galaxy properties from imaging data alone.

1. Introduction

This paper investigates a novel combination of texture descriptors and applies them for automated analysis of galaxy images. We follow the line of filter-based approaches [25, 31, 32] to texture analysis. Specifically, we focus on derivative filters. We construct differential invariants from these filters and agglomerate this information in histogram representations [23]. Descriptors such as SIFT, HoG, and DAISY [26, 12, 30] capture the local structure in images using first order differential structure in the form of gradient orientation histograms. We propose to extend these descriptors by a representation of the second order differential structure. To this end, we suggest using the shape index and the accompanying curvedness measure [21] as the basis for our descriptor, since they provide a summary of the second order structure. The novelty of our approach lies in using localized shape index histograms combined with gradient orientation histrograms both measured at multiple scales. For texture analysis, adding this higher order information will in some applications be necessary in order to improve the discriminative performance of texture representations—and quantifying physical properties of galaxies from imaging data is such an application.

Galactic structure (i.e. how the mass is generally distributed within galaxies) and morphology (i.e. how that mass is arranged on smaller scales) are important diagnostics of the formation and evolutionary mechanisms and timescales for galaxies. It is well known that this structure is correlated with other physical properties of the galaxies such as star-formation rate and dust content (e.g. [7]). However, the means to formalize these relationships are yet to be realized. Extremely large galaxy surveys from the ground, such as the SDSS, have compiled vast, homogeneous imaging of millions of galaxies. Furthermore, ever since the launch of the Hubble Space Telescope (HST) and the advent of adaptive-optics (AO) on large aperture ground-based telescopes enabling high physical-resolution images of galaxies, the study of galaxy structure and morphology has entered a data-rich era.

Galaxies are made of stars, gas and dust. Each of these components emits light over different wavelength ranges and with different intensities. To use the observed light, for example, to determine the mass of stars or the rate at which new stars are being formed, we need to be able to disentangle the various luminous contributions. To do so, astronomers build models of the emission for each source. Gas will primarily emit in emission lines, which appear at a set of discrete wavelengths associated with the emitting element. These emission lines can only be observed spectroscopically and give the most direct measurement of the rate at which new stars are being formed (SFR). Stars, on the other hand, emit continuum radiation over a large range of wavelengths. We can use models of populations of stars as a function of time to extract the mass and age of the stars

in a galaxy. These models can be used for spectroscopy as well as (broad band) imaging in multiple filters (colors).

The mass and SFR of a galaxy can therefore be (coarsely) measured by comparing a set of models with the shape of the spectral energy distribution traced by multiple filters. The specific star formation rate (sSFR) is simply the current SFR divided by the mass of stars. Usually, even if the SFR is determined from emission lines spectroscopically, the mass is determined from the colors of the galaxy in multi-filter imaging. The dominant approach for estimating sSFR from imaging data alone is based on analysis of the color of the galaxy.

Our current knowledge of galaxies is built on imaging surveys and follow-up spectroscopy. Modern imaging surveys will acquire data in several band-pass filters and can be used to approximate galactic properties. However, better determinations of these quantities require deep spectroscopy covering a significant wavelength baseline. Furthermore, most surveys will only have a single band of high angular-resolution imaging (e.g. from space). In such resolved galaxy images, it is possible to use the structure as a proxy for internal dynamics that would require more time-consuming spectroscopic data to observe. Indeed, many of the future surveys will be imaging-only surveys that will not allow for spectroscopic follow-up observations of the vast majority of the observed galaxies. Therefore, being able to fully exploit the most well-resolved images as proxies for spectroscopic data is highly valuable.

Figure 1 illustrates examples of optical images of galaxy from the subset of the SDSS dataset that is used in this paper. The top row shows well-resolved galaxy images. Notice that the light profile of these galaxies contains intricate texture. This texture is caused by the distribution of stars and gas in the galaxy—an important cue for determining the sSFR. We propose to investigate the predictive power of texture when estimating sSFR from optical images. The bottom row of Fig. 1 illustrates problematic cases for our texture based analysis. These range from noise and nearby stars to faint distant galaxies which are poorly resolved in the images. At first glance, this may seem impossible. After all, making the leap from single-band or a few bands imaging data to spectroscopic quantities is a large jump. However, the properties of galaxies are correlated. We have known since the earliest galaxy surveys, that star-forming galaxies have more internal morphological structure due to dust obscuration and star-forming clumps than quiescent (elliptical) galaxies, which tend to be smoother.

There has been some prior work on automated analysis of optical images of galaxies [14, 9]. Much of this work, however, focuses on classification of galaxies based on morphology (e.g. [4]) and tends to ignore information found in the texture. Furthermore, these approaches have used somewhat standard image features as input. Here we present new image features which we believe can capture heretofore ignored information contained in resolved galaxy images.

Figure 1. Examples of low-redshift galaxies in our subset of the SDSS dataset. We have mapped the gri-bands to the RGB color space (gri $\rightarrow$BGR). The top row shows well-resolved galaxies and the bottom row shows problematic cases for our analysis. These color images are best viewed electronically.

The following section (§ 2) describes the galactic dataset we use in our experiments. The new texture features are introduced in § 3. Section § 4 explains how we perform regression in order to predict sSFR values from our features. The results are presented in § 5 before we discuss their implications and future extensions of our work in § 6.

2. Galaxy data

The primary data used for the current work are a sample of low-redshift galaxies drawn from the SDSS DR7, see Fig. 1. We use the g-, r-, and i-band images covering the wavelengths from 4000–5500, 5500–7000 and 7000–8500 Ångstroms, respectively. This sample is defined as all spectroscopic galaxies within the GAMA DR1 region [13] which also have entries in both the MPA-JHU and NYU-VAGC catalogs [11, 8]. The overlap with GAMA for these ~ 12000 galaxies is of particular interest because that survey will acquire spectroscopy of fainter targets and higher quality imaging (including at different wavelengths) thus allowing us, eventually, to extend our analysis to more galaxies and to longer wavelengths.

The images for our galaxy sample were obtained using the *skyview* software provided by NASA/GSFC. For each galaxy position, as defined in the SDSS DR 7, we downloaded a 100×100 pixel region (covering $39.6'' \times 39.6''$) around that position. These images are not background subtracted and do not include an object segmentation map. We used SExtractor [6] on each image to generate and subtract an estimate of the background and to produce a segmentation map including both the target and neighboring galaxies. We have not applied any additional smoothing to the galaxy pixels at this stage because that is a core part of our following analysis. We however compress the intensity range by applying a logarithmic function of the intensities.

The last step in the pre-processing of the images was to construct a refined and well-defined pixel segmentation mask indicating which pixels belonged to the galaxy of interest in each frame. We used a generalized Petrosian method to build these masks, similar to that presented in [3]. We first rank-order the pixels in the SExtractor segmentation map for the target from bright to faint. At each intensity level we calculate the average intensity brighter than that pixel. When the ratio of the pixel's intensity to that average reach a pre-determined value (the Petrosian η) we set that intensity as the lower limit for a pixel to be included in the following analysis. For some galaxies, even with low η, the resulting number of included pixels may be too small for proper analysis (see below for further details). We note that smoothing the data first and then creating the mask will push more pixels above η and create a more inclusive mask. However, these lower significance pixels will not add to the textural features at small scales because they will be highly correlated in a way determined by the smoothing kernel.

Each band image leads to slightly different masks, not only due to noise but also because some galaxy structure is only visible at certain wavelengths. We construct a combined mask by taking the union of the masks for each band. We use this combined mask for processing all of the bands.

The mask extraction (segmentation) occasionally leads to incorrect masks which includes non-galaxy pixels. In order to remove some of these outliers from the analysis, we apply a threshold on the ratio of galaxy pixels and pixels in the convex hull of the galaxy mask. We discard all images where this ratio is less than 0.7.

The galaxy images were extracted such that each galaxy is in the image center. We discard images from the analysis if the mask processing leads to a mask not overlapping with the image center. This may be caused by a faulty mask extraction that latches onto objects in the vicinity such as nearby stars.

In order to remove noise at the boundary of the produced masks and holes inside these, the masks where processed by applying a morphological closing followed by an opening operation with a disk structure element with radius 1 pixel. Following this the masks have been filtered with a linear Gaussian filter with $\sigma = 0.5$ and filter mask size equal to 3σ. This produces a cleaned galaxy mask with smooth boundaries.

Prior to applying the Gaussian filter, we estimate the Petrosian radius of the galaxy by

$$R_p = \sqrt{\frac{N_{\mathrm{gal}}}{\pi}} \,, \tag{1}$$

where N_{gal} denotes the number of galaxy pixels in the mask. Furthermore, we estimate a fiducial orientation of the galaxy from the binary mask, which we use to make the gradient orientation feature invariant to rotation. This estimation is based on the masks prior to Gaussian filtering. We compute the spatial covariance of the galaxy pixels by

$$\mathbf{C}_{\mathrm{gal}} = \frac{1}{N_{\mathrm{gal}} - 1} \sum_{x_{\mathrm{gal}}} (x_{\mathrm{gal}} - \mu)^T (x_{\mathrm{gal}} - \mu) \,, \tag{2}$$

where $x_{\mathrm{gal}} \in \mathbb{R}^2$ is the position of galaxy pixels in the mask, the sum runs over all galaxy pixels in the mask, and

$$\mu = \frac{1}{N_{\mathrm{gal}}} \sum_{x_{\mathrm{gal}}} x_{\mathrm{gal}} \tag{3}$$

is the mean position of all galaxy pixels. We define the fiducial orientation of the galaxy as the eigenvector corresponding to the largest eigenvalue of the covariance matrix. This direction of most spatial variance in galaxy pixels usually corresponds to the major axis of ellipsoidal shaped galaxies. Since the eigenvector is computed up to a change of sign, we flip the sign of any eigenvector with a negative x-component in order to make the orientation consistent. In case of isotropic galaxies this way of picking a fiducial orientation will lead to a random choice, but as there is no natural orientation in this case, this is acceptable.

We note here that our image analysis does not strongly depend on the precise background level (as long as it does not vary greatly on galaxy scales), the choice of η, or on the absolute flux level in the galaxy pixels themselves. Our image features are dependent solely on the intensity texture within the galaxies—not the specific intensity level. That said, objects for which the number of pixels in the mask is smaller than ~ 100 will have insufficient data to reliably measure histogram based image features. We do, however, not remove such images from our study, which potentially leads to outliers in the analysis.

3. Texture descriptors

Discriminative information in textures may appear on several different scales—this is certainly the case for galaxy images—hence using a multi-scale representation appears to be a necessity when performing analysis of texture images. We use the linear scale-space representation [20, 29], where the scale-space of an image $I : \Omega \mapsto \mathbb{R}$, $\Omega \subset \mathbb{R}^2$ is defined as $L(x, y; \sigma) = (I * G)(x, y; \sigma)$, where $*$ denotes convolution with a Gaussian filter

$$G(x, y; \sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \,. \tag{4}$$

The parameter $\sigma > 0$ is the scale of the representation. In this representation we can compute image derivatives of order n and m by

$$L_{x^n y^m}(x, y; \sigma) = \left(I * \frac{\partial^{(n+m)} G}{\partial x^n \partial y^m}\right)(x, y; \sigma) \,. \tag{5}$$

Image derivatives form the basic components of our descriptors, but we will introduce non-linearity in the features by applying functions of these derivatives.

Common descriptors such as SIFT, HoG and DAISY [26, 12, 30] use first order differential structure in the form of gradient orientation histograms as the basis of the descriptor. In smooth scale space derivatives the gradient orientation may be defined as

$$\theta(x, y; \sigma) = \tan^{-1}\left(\frac{L_y(x, y; \sigma)}{L_x(x, y; \sigma)}\right), \quad (6)$$

and the scale normalized gradient magnitude as

$$M(x, y; \sigma) = \sigma^2 \sqrt{L_x^2(x, y; \sigma) + L_y^2(x, y; \sigma)} \ . \quad (7)$$

We need to perform this scale normalisation in order to be able to compare M across different scales σ [29].

We also add a representation of the second order differential structure—namely the shape index and the accompanying curvedness measure [21]. The shape index is based on the eigenvalues κ_1 and κ_2 of the Hessian matrix of the image function. It is defined as the angle between the vector of the eigenvalues (κ_1, κ_2) and the first axis in this eigenvalue space. In terms of image derivatives we may express the shape index as

$$S(x, y; \sigma) = \frac{2}{\pi} \tan^{-1}\left(\frac{-L_{xx} - L_{yy}}{\sqrt{4L_{xy}^2 + (L_{xx} - L_{yy})^2}}\right). \quad (8)$$

The shape index represents the basic second order shapes with dark blobs ($S = -1$), over saddle points ($S = 0$), to bright blobs ($S = 1$), with valley- and ridge-like structure in between. For the detailed geometric interpretation see [21].

The curvedness is simply defined as the length of the eigenvalue vector (κ_1, κ_2) and expresses how pronounced the second order structure is, similar to the role of the gradient magnitude for the first order structure. In terms of image derivatives the scale normalized curvedness may be defined as

$$C(x, y; \sigma) = \frac{1}{2}\sigma^2 \sqrt{L_{xx}^2 + 2L_{xy}^2 + L_{yy}^2} \ . \quad (9)$$

The shape index is rotational invariant by design, contrary to gradient orientation which depends on the choice of coordinate system.

The exact spatial ordering of the texture is not necessarily important, hence it is common (e.g. [26, 12, 30]) to introduce an agglomeration step such as statistical moments or histograms. Here we choose to use smooth histograms inspired by the concept of locally orderless images [22]. This formulation makes the intrinsic parameters of the histogram representation explicit and provides a more robust estimate compared to the traditional histogram formulation.

We define a smooth histogram as a function of the feature f in question and its magnitude F,

$$H(f_i) = \int F(x, y) A(x, y) B(f_i, x, y; f) \, \mathrm{d}x\mathrm{d}y \ , \quad (10)$$

where f_i denotes the histogram binning variable and will act as the bin center for a specific choice of binning aperture function B. The function A localizes the descriptor to specific parts of the image. We propose to use the Gaussian function of β bin width as smooth bin aperture function for histograms of the shape index $S(x, y; \sigma)$

$$B_{\beta,\sigma}(S_i, x, y; S) = \exp\left(-\frac{(S(x, y; \sigma) - S_i)^2}{2\beta^2}\right). \quad (11)$$

The Gaussian bin aperture is not a good choice for gradient orientation histograms, since it does not incorporate the fact that θ is periodic. A better choice is to use the von Mises density function as aperture function, since this is the extension of the Gaussian distribution to the unit circle. We therefore propose to use the following smooth bin aperture function for the gradient orientation $\theta(x, y; \sigma)$

$$B_{\beta,\sigma}(\theta_i, x, y; \theta) = \exp\left(\frac{1}{\beta}\cos\left(\theta(x, y; \sigma) - \theta_i - \theta_0\right)\right), \quad (12)$$

where θ_0 denotes a fiducial orientation.

As feature magnitude F for shape index we will use the curvedness measure C from (9) and for the gradient orientation we will use the gradient magnitude M from (7). The rationale is that we would like local structure with a large magnitude to count more in the histogram. This also has the effect of reducing noise in the histograms caused by noise in the derivative measurements.

We propose to construct texture features by combining histograms of gradient orientation with histograms of shape index and to measure these histograms at different scales σ. As a concrete discretization of this representation we choose an equidistant binning in the histograms and fix the number of bins to 8 for gradient orientation and to 9 for shape index histogram features. The bin width β is chosen such that with the specific choice of number of bins, we tile and cover the complete range of the feature. Equation (10) weights each data point that is added to the histogram by its feature magnitude and each bin window, thus each point casts a vote in every bin of the histogram.

For our specific application to galaxy images we set θ_0 in the gradient orientation feature to be the fiducial orientation of the galaxy as defined in § 2. Furthermore, we choose the window function A in (10) to be identical to the galaxy mask as outlined in § 2. This localizes the feature to include features from only galaxy pixels. In addition, a histogram at a specific scale σ is always normalized so that the bin counts H sum to one.

Notice that our gradient orientation histogram is similar to SIFT-like descriptors, except that we do not include a spatial pooling step (i.e. we only employ a single histogram for the region of interest).

Choosing measurement scales. Using the scale space representation we can compute features at a range of scales capturing pixel correlations across these scales. Selected scales should cover the range of characteristic scales for the particular galaxy image. The inner scale is given by the pixel scale, but since we want to compute derivatives up to second order we need to be careful with the numerics. By choosing the smallest inner scale to be $\sigma_i = 0.88$ pixels we will have less than 1% numerical error in the estimation of the second order derivatives [29]. This inner scale will measure geometry at near pixel level corresponding to $0.396''$.

We approximate the effective outer scale for a particular galaxy image with the Petrosian radius (1). For isotropic galaxies this will be a good estimate, however, for elongated ellipsoidal galaxies this will be a poor over-estimate. We have opted for the simple heuristics of picking the effective outer scale as a function of the Petrosian radius. Let w be the smallest of the image width and height measured in pixels. We then use the Petrosian radius as outer scale $\sigma_o = R_p$ if $4R_p/w \leq 1$, and otherwise choose σ_o such that $4\sigma_o/w = 1$. In order to avoid artifacts in the computed scale space derivatives introduced by boundary effects, it is common to discard pixels that are close to the boundary. The heuristic ensures at least a one σ_o distance from the galaxy to the image boundary. This definition of the outer scale will measure the geometry at galaxy scale. If $\sigma_i > \sigma_o$, we discard the image from the analysis.

We sample the range of effective scales $[\sigma_i; \alpha\sigma_o]$ in exponentially growing steps. We found experimentally that $\alpha = 0.2$ is a good value for the fraction of the outer scale, which focuses the descriptor on the range of scales where relevant structure occurs in galaxy images. We note that this specific choice is application dependent. We choose to use 8 scale levels in the interest of minimizing the computational effort and at the same time achieving good results.

4. SSFR Prediction Experiments

We use regression to predict specific star formation rate (sSFR) from combinations of the texture descriptors outlined above.

Evaluation. We consider different models and feature combinations to predict the sSFR value for each galaxy image. We perform 10-fold cross validation (CV) on our subset of the SDSS dataset. As measure of the prediction error we report the root mean square error (RMSE) averaged over the 10 CV folds. We also report the standard deviation of the RMSE computed from the RMSE on each fold (when interpreting these values it has to be kept in mind that the CV folds are strictly speaking not fully independent).

Models. Because scatter plots indicated a near linear relation between our features and the sSFR, we consider a standard linear least squares regressor as predictor (*Linear*). To further improve the performance, we employ non-linear regression techniques using the Shark machine learning library [18]. We initially considered Gaussian process regression with radial Gaussian kernels, where the bandwidth parameter of the kernel and the precision of the noise were adapted by grid-search as well as gradient-based optimization of the logarithmic marginal likelihood function (or evidence) [27]. However, because the Gaussian processes did not significantly improve over the linear regression, we apply multi-layer perceptron neural networks (*MLP*). Each MLP has a single hidden layer with 100 units with logistic activation functions and a linear output unit. We add shortcut connections linking the inputs directly to the output unit. The training data of each CV fold was further split into an MLP-training and an MLP-validation set using a 9:1 split ratio. The network was trained starting from small weights by minimizing the squared error on the MLP-training set using the iRProp$^+$ first-order optimization algorithm [19]. The weight configuration with the smallest squared error on the MLP-validation set was considered to be the final hypothesis. This "early stopping" of a training process that increases the complexity starting from an (almost) linear model typically fosters good generalizing hypothesis (note that the actual number of hidden units is of lesser importance if chosen large enough, see [5]). In the following, we only report the linear regressor and MLP results.

As a baseline, we use the constant model predicting the sSFR value to be the average sSFR value of the training set (later referred to as *Average*).

We also include a color-based model of the sSFR which was provided together with the SDSS dataset (*Color*). The method is based on the approach described in [16, 15, 28, 10], which employs a physical model of the relations between sSFR and spectrum of a galaxy.

Finally, we augment the color-based physical model by our texture features. This is done by fitting the residuals of *Color*. We refer to the resulting *additive models* [17] as either *Linear-AM* or *MLP-AM* depending on whether linear regression or our neural network approach was used.

Features. As input features, we consider gradient orientation (*GO*) and shape index (*SI*) features as well as their combination (referred to as *All*). Each feature consists of histograms at 8 scale levels.

Furthermore, for reference we include the best results achieved using a feature set consisting of histograms of filter responses for second order directional derivatives and the Laplacian (*2nd*), i.e. the filters used in [31]. These features were implemented using the smooth histograms defined by (10)-(11), and computed at multiple scales using the same choices as for our features.

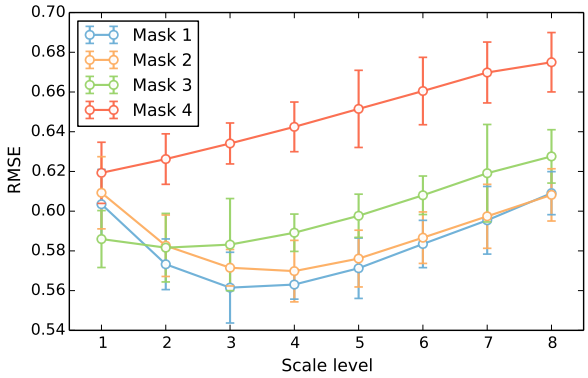


Figure 2. Plot of RMSE (error bars indicate 1 standard deviation of the CV error) of Linear *gri* (SI) across the 8 scale levels for the four masks. Notice for masks 1–2 the curve has a dip indicating that for single scale features an optimal scale exists.

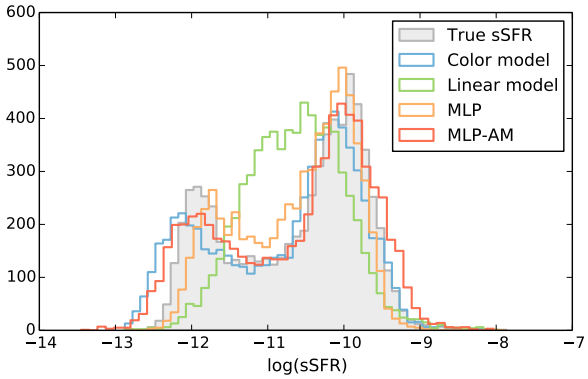


Figure 3. Plot of the distributions of predicted sSFR values for different predictors and the ground truth for mask 1, using the *gri* and shape index (SI) features. It is seen that all models but the linear recover the bimodal sSFR distribution.

We use 4 different mask sizes in decreasing size with mask 4 being the smallest. The amount of galaxy images that passes all inclusion criteria outlined in § 3 for all masks can be found in Table 1.

5. Results and Discussion

Table 1 summarizes our results for different combinations of features extracted from either a single band (g, r, and i) or all bands (gri) and different regressors. The additive models (AM) yield more accurate predictions (2 standard deviations better) than the standard color-based predictor. Thus, the texture features provide information orthogonal to the color model.

Even in single bands the texture information is correlated with the sSFR value, see the Linear and MLP (All) results. Notice that we obtain slightly better accuracies in the g-band. However, the best texture-only results are obtained on the combined gri-bands.

Using gradient orientation features alone does not provide enough information in this particular application. Instead we need to include the shape index feature or use the shape index feature alone. We only include results for the Linear gri predictor, but the tendency is the same for the single bands and the MLP predictor. This is consistent with similar observations made in [24], in which it is argued that increasing differential order of the features can be beneficial for discriminability. The results on the second order features gri (2nd) are comparable to the (all) and (SI) results for mask 1 but with an increased variance, and for masks 2–3 these features are inferior to the shape index (SI) results.

Fig. 2 show the RMSE of the linear regressor based on shape index (SI) features using single scale levels applied to the combined gri features. Remember that, due to our scale range selection procedure (§ 3) for each image the exact scale used at each scale level will vary as a function of the galaxy size. Notice for masks 1–2 the curve has a dip, indicating that for single scale features an optimal scale exists. However the results of Table 1 show that by combining information at several scales simultaneously we are able to obtain better predictions than with a single scale.

The reason for the generally poor results on mask 4 is that these masks tend to only include the galaxy nuclei which usually appears as a bright saturated blob of light. Our texture features does therefore not provide much information at this part of the galaxy.

Our results also indicate that a linear model actually does a good job of fitting the data, but we do get a slight improvement by introducing the non-linear MLP.

To provide some additional insight Fig. 3 show histograms of the spectroscopic sSFR values together with the results of the predictors Linear gri (SI), MLP gri (SI), and MLP-AM gri (SI). All predictors but the linear are able to recover the two known classes of star-forming and quiescent galaxies seen by the two modes in the histograms. Notice how the color-based predictor systematically underestimates the sSFR value (seen by the shift of the histogram to the left) and that the MLP has a tendency to push the modes towards the mean of the dataset. It is evident that the MLP does a better job at recovering the true sSFR distribution than the linear predictor. It can be nicely seen how fitting the residual (MLP-AM) corrects the Color model.

6. Conclusions

We propose to combine gradient orientation and shape index histograms measured at several scales to describe image texture. SIFT-like descriptors include a spatial pooling

Table 1. Summary of our results for different model-feature pairs applied to either single bands (g, r, and i) or all bands (gri) using four different masks (in decreasing size). The results are based on 6880 images passing the inclusion criteria. The numbers in the table indicate RMSE and cross validation standard deviation. *Average* refers to predicting the training data mean (i.e. an estimator of the data variance) and *Color* is the current state-of-the-art physical model (see § 4). *Linear* and *MLP* denote linear and non-linear regression. *Linear-AM* and *MLP-AM* are the additive models combining *Color* with *Linear* and *MLP*, respectively. Gradient orientation (*GO*) and shape index (*SI*) features as well as their combination *All* and second order features (*2nd*) are considered. For more results see the supplementary material.

Method	**Band (features)**		**Mask 1**	**Mask 2**	**Mask 3**	**Mask 4**
Average			0.88 ± 0.02	0.88 ± 0.01	0.88 ± 0.01	0.88 ± 0.01
Color			0.33 ± 0.01	0.33 ± 0.02	0.33 ± 0.02	0.33 ± 0.02
Linear	*g*	(all)	0.61 ± 0.01	0.62 ± 0.02	0.62 ± 0.01	0.65 ± 0.01
	r	(all)	0.65 ± 0.02	0.63 ± 0.02	0.63 ± 0.01	0.67 ± 0.02
	i	(all)	0.65 ± 0.02	0.64 ± 0.02	0.64 ± 0.02	0.67 ± 0.01
	gri	(all)	0.53 ± 0.02	0.54 ± 0.02	0.55 ± 0.02	0.59 ± 0.02
Linear	*gri*	(SI)	0.53 ± 0.02	0.54 ± 0.02	0.55 ± 0.02	0.59 ± 0.01
Linear	*gri*	(GO)	0.81 ± 0.02	0.83 ± 0.01	0.84 ± 0.01	0.85 ± 0.02
Linear	*gri*	(2nd)	0.53 ± 0.03	0.57 ± 0.05	0.68 ± 0.31	0.64 ± 0.05
MLP	*g*	(all)	0.55 ± 0.01	0.57 ± 0.02	0.58 ± 0.02	0.61 ± 0.01
	r	(all)	0.61 ± 0.02	0.59 ± 0.02	0.61 ± 0.02	0.63 ± 0.01
	i	(all)	0.61 ± 0.02	0.60 ± 0.02	0.61 ± 0.01	0.64 ± 0.02
	gri	(all)	0.49 ± 0.02	0.50 ± 0.01	0.52 ± 0.01	0.55 ± 0.02
MLP	*gri*	(SI)	0.50 ± 0.02	0.50 ± 0.01	0.52 ± 0.01	0.56 ± 0.01
Linear-AM	*gri*	(SI)	0.29 ± 0.02	0.29 ± 0.01	0.29 ± 0.02	0.29 ± 0.01
MLP-AM	*gri*	(SI)	0.29 ± 0.02	0.29 ± 0.02	0.29 ± 0.02	0.29 ± 0.02

step collecting information from a grid of histograms tiling the region of interest (ROI). This allows SIFT descriptors to some extend code spatial structure in the ROI beyond first order differential structure. Our gradient orientation feature can be thought of as a single histogram SIFT descriptor. Contrary to general SIFT-like descriptors, we have the luxury of having a segmentation of the object of interest. Instead of applying a spatial pooling step we choose to increase the differential order.

The descriptor introduced in this paper is tuned towards the specific application, predicting the specific star-formation rate (sSFR) from galaxy images, by confining the descriptor to only include information from the galaxy pixels mask. Based on the mask we fix the outer scale used in the scale-space as well as the dominating orientation used in the gradient orientation histogram. However, the descriptor can easily be reconfigured to be constrained to a local image patch and even be extended to a collection of histograms extracted from a spatial pooling scheme such as used in descriptors such as SIFT, HoG and DAISY [26, 12, 30]. The dominating orientation may be estimated following the same approach as in SIFT. Fixing the scale range is application dependent and requires an analysis of the concrete problem under consideration.

The power of the new descriptor is demonstrated in the application of predicting sSFR from imaging data. We obtain good results when using the texture features alone. By combining the color-based physical model with texture information, we outperform the state-of-the-art for sSFR prediction.

The success of the shape index feature can be explained by realizing that what distinguishes a quiescent galaxy from a star-forming one is the distribution of stars, gas, and dust. This leads to the presence or absence of blob-like structures, as well as the occurrance of ridge-like structures caused by spiral arms and stripe patterns formed by the distribution of gas and dust—the shape index is tuned to this type of second order structure.

A current limitation of the approach is that we extract features independently from each band image ignoring the natural correlation across bands. A future extension would be to extract color descriptors by extending the shape index descriptor to be based on the Hessian matrix of the 2D intensity manifold embedded in the spatio-color space. This strategy would also be readily applicable on other types of color image data.

One of the challenges for computer vision and machine learning in astrophysics is to take models and knowledge

gained from one training set (i.e. a particular survey) and apply it to data taken using different telescopes, instruments and techniques. For our current efforts, the primary difference will be the absence of the spectroscopic ground truth for current and future galaxy surveys. Many of the largest planned surveys are indeed imaging-only and while some spectroscopic follow-up will be done, it will be impossible to obtain complete spectroscopic coverage of the more numerous (and often fainter) galaxies being imaged. Against this background, this study is the first step towards enabling the quantification of physical galaxy properties from imaging data alone. We expect that this mapping of galaxy appearance and properties will prove extremely useful when applied to future large scale imaging-only surveys such as the Large Synoptic Survey Telescope (LSST).

Acknowledgements

The authors thank the SDSS [2] and GAMA [1] for making the galaxy data available and gratefully acknowledge support from The Danish Council for Independent Research (FNU 12-125149).

References

[1] http://www.gama-survey.org.

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

[3] R. G. Abraham, P. Nair, P. J. McCarthy, et al. The Gemini Deep Deep Survey. VIII. When Did Early-Type Galaxies Form? *Astrophys. J.*, 669:184–201, 2007.

[4] M. Banerji, O. Lahav, C. J. Lintott, et al. Galaxy zoo: Reproducing galaxy mophologies via machine learning. *MNRAS*, 406:342–353, 2010.

[5] P. L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE T Inform Theory*, 44(2):525–536, 1998.

[6] E. Bertin and S. Arnouts. SExtractor: Software for source extraction. *Astron. Astrophys. Supp.*, 117:393–404, 1996.

[7] M. R. Blanton and J. Moustakas. Physical Properties and Environments of Nearby Galaxies. *Ann. Rev. Astron. Astrophys.*, 47:159, 2009.

[8] M. R. Blanton, D. J. Schlegel, M. A. Strauss, et al. New York University Value-Added Galaxy Catalog: A Galaxy Catalog Based on New Public Surveys. *Astron. J.*, 129:2562, 2005.

[9] S. Bridle, S. T. Balan, M. Bethge, et al. Results of the GREAT08 challenge: An image analysis competition for cosmological lensing. *MNRAS*, 405(3):2044–2061, 2010.

[10] J. Brinchmann. Private communication.

[11] J. Brinchmann, S. Charlot, S. D. M. White, et al. The physical properties of star-forming galaxies in the low-redshift Universe. *MNRAS*, 351:1151, 2004.

[12] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893, 2005.

[13] S. P. Driver, D. T. Hill, L. S. Kelvin, et al. Galaxy and Mass Assembly (GAMA): survey diagnostics and core data release. *MNRAS*, 413:971, 2011.

[14] G. Fasano, E. Vanzella, A. Dressler, et al. Morphology of galaxies in the WINGS clusters. *MNRAS*, 420(2):926–948, 2012.

[15] A. Gallazzi, J. Brinchmann, S. Charlot, et al. A census of metals and baryons in stars in the local Universe. *MNRAS*, 383:1439–1458, 2008.

[16] A. Gallazzi, S. Charlot, J. Brinchmann, et al. The ages and metallicities of galaxies in the local universe. *MNRAS*, 362:41–58, 2005.

[17] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman & Hall, 1990.

[18] C. Igel, T. Glasmachers, and V. Heidrich-Meisner. Shark. *JMLR*, 9:993–996, 2008.

[19] C. Igel and M. Hüsken. Empirical evaluation of the improved Rprop learning algorithm. *Neurocomputing*, 50(C):105–123, 2003.

[20] J. J. Koenderink. The structure of images. *Biol. Cybern.*, 50:363–370, 1984.

[21] J. J. Koenderink and A. J. van Doorn. Surface shape and curvature scale. *Image Vision Comput.*, 10(8):557–564, 1992.

[22] J. J. Koenderink and A. J. van Doorn. The structure of locally orderless images. *IJCV*, 31(2/3):159–168, 1999.

[23] J. J. Koenderink and A. J. van Doorn. Local structure of Gaussian texture. *IEICE Trans Inf Syst*, E86-D(7):1165–1171, 2003.

[24] A. B. L. Larsen, S. Darkner, A. L. Dahl, and K. S. Pedersen. Jet-based local image descriptors. In *ECCV 2012*, volume LNCS 7574, pages 638–650. Springer, 2012.

[25] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43(1):29–44, 2001.

[26] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[27] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

[28] S. Salim, R. M. Rich, S. Charlot, et al. UV Star Formation Rates in the Local Universe. *Astrophys. J. Suppl. Ser.*, 173:267–292, 2007.

[29] B. M. ter Haar Romeny. *Front-End Vision and Multi-Scale Image Analysis*. Kluwer Academic Publishers, 2003.

[30] E. Tola, V. Lepetit, and P. Fua. DAISY: An efficient dense descriptor applied to wide-baseline stereo. *IEEE TPAMI*, 32(5):815–830, 2009.

[31] M. Varma and A. Zisserman. A statistical approach to texture classification from single images. *IJCV*, 62(1/2):61–81, 2005.

[32] S. C. Zhu, Y. N. Wu, and D. Mumford. Filters, random fields, and maximum entropy (frame): Towards a unified theory for texture modeling. *IJCV*, 27(2):107–126, 1998.

8

Sacrificing information for the greater good: how to select photometric bands for optimal accuracy

Bibliographic reference

K. Stensbo-Smidt, F. Gieseke, C. Igel, A. Zirm, and K. Steenstrup Pedersen. Sacrificing information for the greater good: how to select photometric bands for optimal accuracy. *Monthly Notices of the Royal Astronomical Society*, 464(3):2577–2596, 2017.

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **464,** 2577–2596 (2017) doi:10.1093/mnras/stw2476
Advance Access publication 2016 September 30

Sacrificing information for the greater good: how to select photometric bands for optimal accuracy

Kristoffer Stensbo-Smidt,[1⋆] Fabian Gieseke,[1] Christian Igel,[1,2⋆] Andrew Zirm[3] and Kim Steenstrup Pedersen[1,2⋆]

[1]*Department of Computer Science, University of Copenhagen, Universitetsparken 5, DK-2100 Copenhagen, Denmark*
[2]*Space Science Center, University of Copenhagen, Juliane Maries Vej 30, DK-2100 Copenhagen, Denmark*
[3]*Dark Cosmology Centre, Niels Bohr Institute, University of Copenhagen, Juliane Maries Vej 30, DK-2100 Copenhagen, Denmark*

Accepted 2016 September 28. Received 2016 September 23; in original form 2015 November 9

ABSTRACT

Large-scale surveys make huge amounts of photometric data available. Because of the sheer amount of objects, spectral data cannot be obtained for all of them. Therefore, it is important to devise techniques for reliably estimating physical properties of objects from photometric information alone. These estimates are needed to automatically identify interesting objects worth a follow-up investigation as well as to produce the required data for a statistical analysis of the space covered by a survey. We argue that machine learning techniques are suitable to compute these estimates accurately and efficiently. This study promotes a feature selection algorithm, which selects the most informative magnitudes and colours for a given task of estimating physical quantities from photometric data alone. Using k-nearest neighbours regression, a well-known non-parametric machine learning method, we show that using the found features significantly increases the accuracy of the estimations compared to using standard features and standard methods. We illustrate the usefulness of the approach by estimating specific star formation rates (sSFRs) and redshifts (photo-z's) using only the broad-band photometry from the Sloan Digital Sky Survey (SDSS). For estimating sSFRs, we demonstrate that our method produces better estimates than traditional spectral energy distribution fitting. For estimating photo-z's, we show that our method produces more accurate photo-z's than the method employed by SDSS. The study highlights the general importance of performing proper model selection to improve the results of machine learning systems and how feature selection can provide insights into the predictive relevance of particular input features.

Key words: methods: data analysis – methods: statistical – techniques: photometric – galaxies: distances and redshifts – galaxies: star formation – galaxies: statistics.

1 INTRODUCTION

High-resolution spectroscopic data contain a wealth of information about astrophysical objects. Analyses relying on spectroscopy suffer, however, from small sample sizes. Photometric surveys have the potential to overcome this limitation, but are limited in terms of the amount of information that can be extracted for each astrophysical object. Due to the abundance of data currently available, and especially with the surveys commencing within the next decade, methods are required that can automatically extract relevant information from the broad-band images of these surveys. Our goal is to reliably, efficiently, and accurately estimate properties of objects from photometric data, for example, for quickly identifying interesting objects worth a follow-up investigation or for conducting large-scale statistical analyses. In this study, we apply a method for selecting the most informative colours and bands for photometric estimations. We illustrate its potential by estimating specific star formation rates (sSFRs) and photometric redshifts (photo-z's) from available Sloan Digital Sky Survey (SDSS) data, but the method can readily be applied to other quantities and surveys.

1.1 Star formation rates

An ongoing quest in cosmology is the understanding of galaxy formation and evolution. A crucial part here is to understand the star formation history (SFH) of the individual galaxies as well as the Universe as a whole. Major open questions include which processes trigger star formation and, equally important, quench it. Data from large surveys, such as the SDSS (York et al. 2000), have shown a peculiar bimodality in the star formation rates (SFRs) of

⋆ E-mail: k.stensbo@di.ku.dk (KS-S); igel@di.ku.dk (CI); kimstp@di.ku.dk (KSP)

galaxies (Kauffmann et al. 2003). The bimodality points to a scenario where star formation is quenched, but the responsible mechanism is far from understood. Current results indicate that the quenching time-scale varies significantly with galaxy mass (Wetzel, Tinker & Conroy 2012; Wetzel et al. 2013; Wheeler et al. 2014) and redshift (Balogh et al. 2016), suggesting that different processes are in play at different times and masses (Fillingham et al. 2015; Wetzel, Tollerud & Weisz 2015). To uncover these processes, it is natural to turn to the statistical properties of a large number of galaxies in order to look for correlations between SFRs and other physical properties.

The most common way to estimate the recent SFR of a galaxy is to use a number of observational tracers. These tracers often rely on observations of single or multiple emission lines, with the Hα emission line being among the most popular (Kennicutt & Evans 2012). A main limitation is that they usually require high-quality spectra. Other types of observational tracers are derived from broad-band observations. Based on these, one may estimate SFRs using conversion factors to convert from flux over a given wavelength interval (Kennicutt 1998; Kennicutt & Evans 2012).

Unfortunately, SFRs derived from different observational tracers are not always consistent. There are many reasons for these inconsistencies. For instance, different observational tracers are sensitive to different types of star formation, galaxy populations, redshifts, etc. There are also the problems of correcting observations (e.g. for dust) and how to define the boundary of a galaxy when integrating the light from it. Davies et al. (2016) found various degrees of inconsistencies between different SFR indicators for a selection of spiral galaxies. They attempted to correct these inconsistencies by recalibrating the methods using a linear relationship between luminosity and SFR.

One may also estimate SFRs using spectral energy distribution (SED) fitting, which relies on a library of template spectra generated by stellar population synthesis models (for recent reviews, see Walcher et al. 2011; Conroy 2013). In the most basic version of this method, an observed galaxy spectrum is compared to every template spectrum, the closest match is chosen, and the template's physical properties adopted (e.g. Charlot et al. 2002; Brinchmann et al. 2004).

SED fitting is often considered a less precise way to estimate SFRs than relying on observational tracers (Walcher et al. 2011), but it can be done with libraries such as CIGALE (Noll et al. 2009) and MAGPHYS (da Cunha, Charlot & Elbaz 2008). SED fitting also allows us to estimate the SFR from broad-band photometry, where observational tracers have more limited use (Maraston et al. 2010).

More direct estimations of SFRs and sSFRs from broad-band photometry have also been investigated (e.g. Williams et al. 2009; Arnouts et al. 2013), though there are still significant discrepancies between these estimated quantities and those obtained from more reliable methods.

1.2 Photometric redshifts

Spectroscopic surveys provide highly accurate redshifts of galaxies, enabling a detailed 3D view of galaxy distribution in the Universe, but they are both expensive and time consuming. Photometric surveys, on the other hand, can cover a much larger area of the sky in less time, and can usually go below the spectroscopic flux limit. They therefore provide a significantly more complete, and thus less biased, sample of galaxies, which is a notable advantage over spectroscopic surveys. Photometric surveys, however, struggle with reduced accuracy in the galaxy positions along the line of sight. Despite this problem, the larger galaxy sample sizes are useful for numerous cosmological applications, such as obtaining constraints on cosmological parameters (e.g. Padmanabhan et al. 2007; Carnero et al. 2012; Ho et al. 2012). These applications rely on photometric redshifts (photo-z's) calculated from broad-band photometry. Naturally, increasing the accuracy of photo-z's is of great importance.

A vast amount of methods have been developed to estimate photo-z's (see e.g. Hildebrandt et al. 2010; Abdalla et al. 2011, for recent comparisons). Broadly speaking, photo-z estimation methods can be classified as either template-based or empirical methods. Template-based methods use SED fitting in the same way as for SFR estimation; they match the observed colours or magnitudes to those of a large library of synthetic template spectra (e.g. Benítez 2000; Bolzonella, Miralles & Pelló 2000; Ilbert et al. 2006; Brammer, van Dokkum & Coppi 2008).

Empirical methods train algorithms to estimate photo-z's from colours or magnitudes. The algorithms are calibrated to fit the task at hand using a training data set with spectroscopically derived redshifts.

A wide range of empirical methods have been developed, and most fall into the categories of either tuning the colour–z relation or machine learning. The machine learning category is highly diverse, with techniques such as artificial neural networks (Collister & Lahav 2004), self-organizing maps (Geach 2012), random forests (Carrasco Kind & Brunner 2013), and Gaussian processes (Almosallam et al. 2016) having been used for photo-z estimation. These techniques generally outperform template-based methods for photo-z estimation, as machine learning methods are able to adapt to the highly non-linear relation between colours and redshift. For recent reviews of the performances of various photo-z estimation methods, see Dahlen et al. (2013) and Sánchez et al. (2014).

1.3 Increasing the information from photometric measurements

SED fitting is a common method for both photo-z and SFR estimation. Advantages of this method include the ability to get the full SFH (limited by the detail level of the template library) of a galaxy as well as constraints on its redshift, environment, etc. The restrictions lie in the generation of the template spectra, with computational power and understanding of stellar evolution being the main limiting factors.

The main computational limitation is the enormous amount of free parameters that can be tweaked in the generation of a single spectrum. Because of this, and limited physical knowledge about stellar evolution, it is still a great challenge to generate appropriate template spectra (e.g. Pacifici et al. 2015; Smith & Hayward 2015). A brute-force way of calculating templates for a chosen grid of parameters quickly becomes infeasible. The amount of degeneracies between the evolutionary states of different single stellar populations (SSPs) also limits this approach.

A number of ways to reduce the amount of necessary template spectra with minimum information loss have been explored. In particular, machine learning methods have been used to interpolate between template spectra to allow for a sparser grid to be sampled (e.g. Tsalmantza et al. 2007). Active learning was explored by Solorio et al. (2005), where the computer automatically generates new template spectra if no close match is found in the data set. This automatically refines the template grid in regions that have actual observations. A different approach was taken by Richards et al. (2009), who used diffusion K-means to tackle the problem of

choosing which SSPs make up a galaxy spectrum, by finding an appropriate basis from a large set of SSP spectra. In the same spirit, Chen et al. (2009) used a principal component analysis to estimate sSFRs from obtained eigenspectra.

While spectroscopy is superior in terms of information content, photometry excels in terms of coverage. Using a machine learning approach to estimate parameters can give us the best of both worlds. The algorithm can be trained on galaxies with accurate parameters determined from high-resolution spectra and then be used to estimate the same parameters of other galaxies from broad-band photometry only. This avoids the problem of generating template spectra from models that may suffer from various restrictions and approximations. However, just as template-based methods require the parameter space to be densely sampled in order to provide good parameter estimations, machine learning methods require training data that represent the entire population. If such are not available, the methods may lead to biased estimates. Machine learning methods can also achieve significantly lower computational complexity compared to SED fitting, depending on the level of detail wanted, which will become increasingly important in the near future, when new photometric surveys start producing data at an unprecedented rate.

Using highly detailed data can, however, lead to a decrease in accuracy. This counterintuitive phenomenon occurs for both template methods as well as machine learning methods, and can be attributed to the fact that if a dimension contributes only (or even just some) noise, it will decrease the overall signal-to-noise ratio (S/N).

Selecting only the most informative dimensions of the data can therefore lead to higher accuracy, even if it requires removing somewhat informative dimensions, as the lower dimensionality of the data can result in a higher S/N.

In the machine learning literature, the dimensions of a data point are referred to as *features*. Thus, the task of choosing the most informative dimensions is called *feature selection*. Feature selection has already been investigated in an astrophysical context. Among the most used feature selection algorithms are random forests, which produce feature ranking as part of the algorithm. They have been used in a number of studies, for example, D'Isanto et al. (2016) and Rimoldini et al. (2012). Random forests are not the only way to select features, and Graham et al. (2013) tested five different feature selection strategies for classifying stars. Hoyle et al. (2015) showed how adding the most informative features to the standard set of colours and magnitudes significantly increased the accuracy for photo-z estimation.

It is important to realize that the concept of most informative features is not a universal one; the most informative features for one algorithm may be different from those of another. That depends on how specifically the algorithm uses the features, for example, some algorithms may be sensitive to scaling of the features, while others may not. And just as the most informative features vary from algorithm to algorithm, so will they vary from task to task. For example, whereas observed UV radiation may contain a lot of information regarding star formation in the nearby Universe, it may not be that informative for detecting, say, brown dwarfs.

In this paper, we show that we can obtain a significantly greater accuracy of estimated photo-z's and sSFRs, using only SDSS *ugriz* photometry, by applying a machine learning method rather than relying on spectral modelling of the photometry. Our approach is similar to that of Stensbo-Smidt et al. (2013), but here we show that the accuracy can be further increased by performing a feature selection, selecting the most informative features among all measured SDSS magnitudes and colours.

Specifically, we use k-nearest neighbours (k-NN) regression, which is an intuitive method well known in machine learning and to some extent also in astronomical communities (see e.g. Li, Zhang & Zhao 2008; Polsterer, Zinn & Gieseke 2013; Polsterer et al. 2014; Kremer et al. 2015; Kügler, Polsterer & Hoecker 2015). Of the more prominent uses of k-NN in astronomy is the estimation of photo-z's in SDSS (Abazajian et al. 2009).

By using k-NN, we can automatically learn a mapping from magnitudes and colours of galaxies to their parameters derived from reliable indicators, thereby allowing accurate photometric estimates without high-resolution spectra. The reliable parameters can be estimated using any method deemed appropriate for each individual galaxy, effectively taking advantage of multiple indicators, as explored by Wuyts et al. (2011, 2013) for SFRs. A significant advantage of k-NN over other methods is that it naturally adapts to the local, potentially high-dimensional structure of the data, and can thus model highly non-linear behaviour without problems. Another virtue of k-NN is its simplicity, which makes it easy to see how data are used and compared within the algorithm.

Selecting the most informative features can, in theory, be done by trying all possible feature combinations. As the number of combinations grows exponentially with the number of features, this quickly becomes unfeasible, and one has to resort to clever selection strategies. Here, we use *forward feature selection* to determine the most informative features (see Section 2.3 for details). Forward feature selection was used by Xu et al. (2013) to examine which halo properties contained most information about the number of galaxies. In this paper, we use it to improve the estimation of photo-z's and photometric sSFRs, which illustrate the method's general usefulness.

The remainder of this paper is organized as follows: in Section 2, we describe the k-NN algorithm and the algorithm we use to select the most informative colours. Section 3 describes the data we are using and details our experimental set-up. In Section 4, we provide results of our experiments and an analysis of these. We end with a discussion and a summary of our conclusions in Section 5.

2 METHODS

The goal of this study is to test the efficiency of machine learning techniques, in particular feature selection, when estimating physical quantities of galaxies. We suggest using the selected features directly in regression methods rather than in connection with physical models, such as population synthesis models. There are two fundamental ways of doing regression: parametric and non-parametric. In the parametric case, data are assumed to follow a function $f(x)$ with known form but unknown parameters. It is usually fairly easy to estimate these parameters by fitting, but this advantage comes at a cost: by choosing a particular functional form of $f(x)$, we have made assumptions about the underlying structure of the data. If these assumptions are not absolutely correct, we will not be able to achieve optimal estimation performance (James et al. 2013). This is where non-parametric methods have an advantage, as they do not make any assumptions about the structure of the data, but adapt to it.

2.1 k-NN regression

We employ one of the simplest non-parametric methods, namely *k-NN regression* (Altman 1992; Hastie, Tibshirani & Friedman 2009; James et al. 2013). Assume that we are given a data set $\mathcal{S} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\} \subset \mathbb{R}^D \times \mathbb{R}$ consisting of D-dimensional

data points $\boldsymbol{x}_i$ with associated output values y_i. For instance, each data point could represent a galaxy with $D = 2$ colour values (e.g. $B - V$ and $U - B$) and the output value y_i could be the sSFR that one is interested in estimating. The components of $\boldsymbol{x}_i$ (which, in this example, would be the colours) are called *features*. Now, we employ machine learning to infer from S a general rule of how to predict the (unknown) output value y' given some new data point $\boldsymbol{x}'$. The k-NN method does this by simply finding the k closest data points with known output values, and then taking the average of these values, i.e.

$$y' = \frac{1}{k} \sum_{i \in \mathcal{N}_k} y_i, \tag{1}$$

where $\mathcal{N}_k$ is the set of the k nearest data points in $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ w.r.t. the new sample $\boldsymbol{x}'$. The 'closeness' between samples is defined via a metric d. That is, $\mathcal{N}_k = \mathcal{N}_{k-1} \cup \text{argmin}_{(x,y) \in S \setminus \mathcal{N}_{k-1}} \text{d}(\boldsymbol{x}, \boldsymbol{x}')$ for positive integers k and $\mathcal{N}_0 = \emptyset$, where argmin breaks ties at random. We use the Euclidean metric $d(\boldsymbol{x}, \boldsymbol{z}) = \sqrt{\sum_{i=1}^{D} (x_i - z_i)^2}$ for $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^D$, though any metric can be chosen. The Euclidean distance is the most common choice in the literature, but it is perfectly possible that another metric would perform better. One can also attempt to learn the metric from the data as done by, e.g. Weinberger & Saul (2009). To keep things simple, however, we stick to the Euclidean metric.

Although the k-NN regression method is simple, it often yields highly accurate predictors. This is especially the case if the amount of training data N is large and/or the feature space dimensionality D is low. While it may seem counterintuitive, adding more features (i.e. dimensions) to the input data may make k-NN perform worse. The performance of nearest neighbours models can deteriorate if D gets too large, in particular when each added dimension contains intrinsic noise. The addition of extra noise with each added dimension may eventually decrease the S/N. This is perhaps most easily recognized if one considers the extreme case of adding a feature, which is pure noise. This can only decrease the performance, and adding more of these pure noise features will eventually drown any signal present in the original features.

Thus, it is important to select the right features, see Section 2.3.

2.1.1 k-*NN and non-linear recalibration*

As mentioned in Section 1.1, there is often disagreement between various observational tracers when estimating SFRs. Davies et al. (2016) recalibrated a number of observational tracers to produce more consistent SFR estimates, using a linear relationship between observed luminosity and derived SFR.

Such a recalibration is not without its own problems, though. First, a 'true' SFR has to be defined as the base that other estimation methods will be calibrated to. Defining such a true SFR is problematic in itself. Secondly, a linear relationship may not be flexible enough to capture the variability in the data.

Our proposed k-NN method can be seen as being similar in spirit to a recalibration. The method cannot infer SFRs without access to a training set, meaning that the SFRs need to be estimated using another method beforehand. Thus, the k-NN method is in fact modelling a potentially very accurate SFR estimator, while having access only to some less informative features. In this paper, for example, accurate sSFRs are derived from spectroscopy, and the k-NN method is modelling these using only photometric information. Since k-NN is a highly non-linear method, it should be able to obtain better estimates than simpler linear methods. Furthermore, because of the non-linearity, we do not need to restrict the estimations to particular subsets of the data, such as spiral galaxies, but can model the entire population simultaneously.

2.1.2 *Dealing with uncertainties*

In its most basic form, the k-NN algorithm does not support the inclusion of uncertainties associated with inputs or outputs, nor does it provide confidence intervals for the estimated quantities beyond calculating the variance of the neighbours' outputs (Altman 1992). There are, however, extensions dealing with these issues.

There are a number of ways uncertainties may influence the results of an analysis. First, there may be uncertainties related to the output values (e.g. sSFRs or photo-z's) of the training data, which need to be propagated to the predicted output. Secondly, there may be uncertainties in the input values (e.g. colours) of both the training data and the new data, which also need to be propagated to the estimated output value.

Propagating uncertainties from known data to the estimate made by k-NN is not a trivial task. Ideally, to estimate the output value of a new datum, its input uncertainties need to be propagated, and one needs to incorporate the uncertainties on both input and output of the training data. A standard Monte Carlo sampling can deal with all these uncertainty issues, but it will quickly get far too computationally expensive.

Assuming Gaussian errors, uncertainty in the output alone can be dealt with in a relatively straightforward manner by using a weighted average, $y' = \sum_{i \in \mathcal{N}_k} w_i y_i / \sum_{i \in \mathcal{N}_k} w_i$, using $w_i = \sigma_i^{-2}$, where σ_i^2 is the variance of y_i. This does not account for the scatter of the inputs, which ideally should mean that more distant neighbours (and their corresponding uncertainties) are weighted less when computing the average. This can be accounted for by including the similarity metric in the weights, or, alternatively, including the uncertainties in the similarity metric as done by, e.g. Polsterer et al. (2013). An additional complication arises due to the uncertainties in the inputs and the choice of number of neighbours, k. With uncertain inputs, the question of which of two neighbours is closer cannot be answered with complete certainty.

Both the question regarding choosing k and that of choosing the proper similarity metric can, however, be addressed with a probabilistic formulation of k-NN (Holmes & Adams 2002; Everson & Fieldsend 2004; Manocha & Girolami 2007), which allows for posterior inference over k and the similarity metric.

Finally, one may simply try to find a heuristic, reasonable estimate of the uncertainty of the new data. This is, for instance, how the photo-z uncertainties in the SDSS data base have been computed (Abazajian et al. 2009). Here, a hyperplane was fitted to the nearest 100 neighbours in colour space, and the mean deviations of the redshifts from this hyperplane were found to be good estimates of the errors.

To our knowledge, there is no accepted way of dealing with all uncertainty issues short of Monte Carlo sampling. In this paper, we have therefore chosen to ignore the question relating to uncertainties, focusing solely on demonstrating the performance gain of combining k-NN and feature selection.

2.2 Choosing the number of neighbours

In most versions of k-NN, including the vanilla version, one must choose the number of neighbours, k, to average over. Increasing k implies that a prediction will be based on the average of many samples, which reduces the variance of the classifier but may increase

its bias (for a discussion of the bias-variance decomposition of the error of k-NN regression, we refer to Hastie et al. 2009). A standard technique for choosing k is cross-validation (CV). In M-fold CV, the available data S is randomly partitioned into M subsets $S_1, \ldots, S_M$ of (almost) equal sizes. Let $S_{\setminus i} = \bigcup_{j=1,\ldots,M \wedge j \neq i} S_j$ denote all data points except those in S_i. For each $i = 1, \ldots, M$, an individual model is built by applying the algorithm to the training data $S_{\setminus i}$. This model is then evaluated using the test data in S_i. The average error is called *CV error* and is a predictor of the generalization performance of the algorithm. To choose k for k-NN using M-fold CV, S is split into M subsets. For each fold $i = 1, \ldots, M$, k-NN models are built and tested using different values for k (say, $k = 1, 3, 5, \ldots$). The k with the lowest CV error is finally selected.

It must be stressed that the data used for model selection must be independent from data for assessing the final performance of a model.

2.3 Informative features

The use of appropriate features is crucial for machine learning. Standard features in astronomy are, for instance, magnitudes or the derived colours. The performance of a model can, however, often be improved by considering additional features or special combinations of features (thus, effectively changing the underlying distance metric d).[1] We employ automatic *feature selection* to pick the most informative features for our regression task.

2.3.1 Feature selection

The goal of feature selection is to reduce the dimensionality of the input space by selecting the most informative features. A direct way to select such informative features is to systematically try various combinations of them and select the subset with the most promising accuracy for the final model (based on a certain evaluation criterion such as CV). In theory, one would like to try every possible combination of features, but in practice this is often infeasible due to the induced exponential runtime. In the literature, different techniques have been proposed to address this issue, such as the idea to maximize the probability of finding the best combination of features. We refer to Guyon & Elisseeff (2003) for an introduction to feature selection.

Standard alternatives to such an exhaustive search are *forward* and *backward feature selection* (Hastie et al. 2009), which aim at selecting informative features in an incremental manner. For the case of forward selection, one starts by selecting the most promising feature by assessing the predictive power of each of the D features. In the second iteration, the first feature is kept and a second one is selected based on the predictive power of both the first *and* the second feature. This process is repeated until the number $\bar{d}$ of desired features is selected. Backward elimination works similarly. However, instead of incrementally adding features, one removes a feature at a time, starting with all D features being selected.

Even forward and backward feature selection are still computationally demanding, but using clever implementations and data structures one may parallelize the procedure. This paper uses a massively parallel matrix-based implementation combining incremental feature selection and nearest neighbour models, recently proposed by Gieseke et al. (2014a). For more details on the implementation, we refer to Appendix A.

[1] For instance, the SDSS pipeline resorts to two different types of magnitudes via the linear model $\texttt{psfMag} - \texttt{cModelMag} > 0.145$ to classify photometric objects as 'galaxy' or 'point-like'.

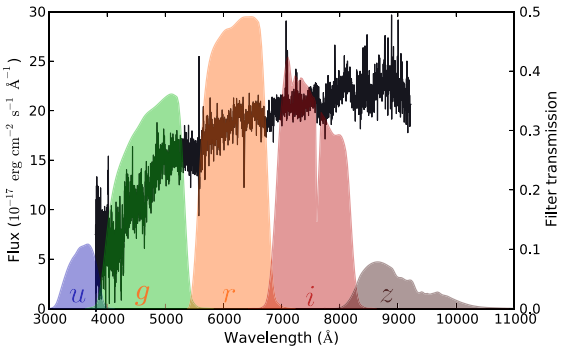


Figure 1. An example spectrum of a galaxy from the SDSS data base (black curve) overlaid by the five bandpass filters of SDSS (Fukugita et al. 1996).

3 EXPERIMENTAL SET-UP

3.1 Data selection

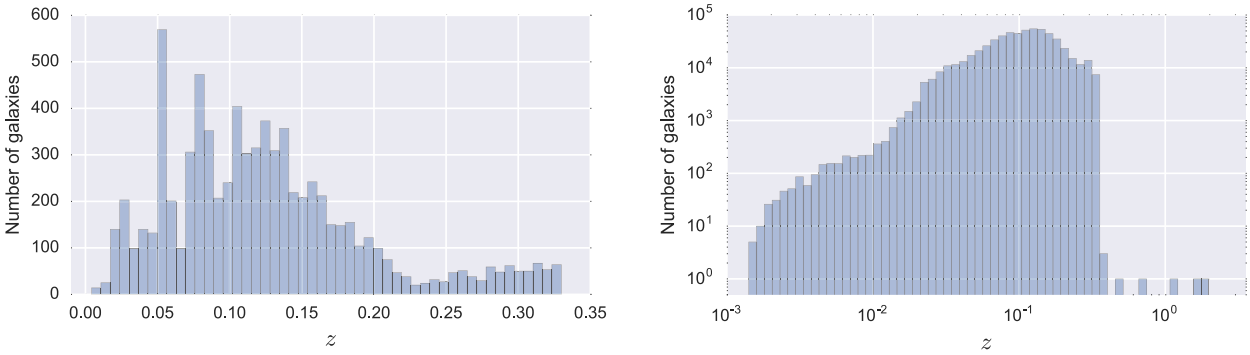
The experiments in this study use photometric data from SDSS (York et al. 2000). The data are a subset of SDSS Data Release 7 (DR7; Abazajian et al. 2009), and consist of `psfMag`, `fiberMag`, `petroMag`, `deVMag`, `expMag`, and `modelMag` magnitudes in the *u, g, r, i,* and *z* bands (see Fig. 1) for each galaxy as well as the galaxy's sSFR and redshift, estimated from spectroscopy. We also include the photometric redshifts estimated by SDSS (Abazajian et al. 2009).

Data are obtained from SDSS CasJobs, using the `SpecPhoto` view, which ensures that objects have clean spectra. sSFRs were taken from Brinchmann et al. (2004).[2] To clean the data, we apply the following constraints.

(i) For sSFRs, we require that the estimation was successful (`flag = 0`), and we remove all duplicate galaxies.
(ii) For redshifts, we require that both spectroscopic and photometric estimations were successful (for spectroscopy, `zWarning = 0`; for photometry, `zErr >= 0`).

A sample of 611 479 galaxies meet the above criteria. For a smaller subset of 7799 low-redshift galaxies ($0.0042 < z < 0.33$) within the selected sample, we additionally have photometric sSFR estimations obtained by a template-based modelling approach described in Section 3.2. No additional selection criteria have been applied to this subset. In particular, no S/N cut has been used in order to highlight the method's robustness to varying noise levels. In this work, we do not make use of any S/N information, though special treatment of low S/N sources can be incorporated in various ways (see discussion in Section 2.1.2). The experiments will be based on two samples of galaxies: the smaller subset and the full sample, excluding the smaller subset (totalling 603 680 galaxies). The redshift distributions of these two samples can be seen in Fig. 2. The smaller subset consists entirely of low-redshift galaxies, where also the majority of the larger sample can be found. The larger sample consists primarily of galaxies below $z \sim 0.5$, with only a few galaxies at higher redshifts.

[2] We used `specsfr_avg` from the data located at http://wwwmpa.mpa-garching.mpg.de/SDSS/DR7/sfrs.html.



(a) Small subset. (b) Entire sample, exluding smaller subset.

Figure 2. Redshift distribution of the two galaxy samples used in the experiments. The entire sample, excluding the smaller subset, additionally contains a single galaxy with $z = -1.93 \times 10^{-4}$, which is not shown in the plot.

3.2 Comparison with other methods

We will compare our results to those of two other methods, one for the sSFR estimations and one for the photo-z estimations.

For the photo-z experiments, we compare our results to the photo-z's available directly through the SDSS data base. These photo-z's have been estimated using a combination of k-NN (Csabai et al. 2007) and a template-based method (Budavári et al. 2000), as described in Abazajian et al. (2009). The k-NN part of the method differs from our approach in that it bases the estimated photo-z on a local hyperplane fitted to the 100 nearest neighbours, instead of just taking the average (and optimizing the number of neighbours), as we do. A potential benefit of using a hyperplane is the ability to extrapolate beyond the redshifts of the training set. Since the vanilla k-NN estimates redshifts by computing the average redshift of the nearest neighbours (in, for instance, colour space), its estimates are restricted to be within the minimum and maximum values of the training set. A linear model (a hyperplane), on the other hand, can extrapolate beyond these, but the quality of the estimates will depend on how well the local neighbourhood is described by the linear model.

For the sSFR experiments, we compare our estimated sSFRs to those obtained by the standard approach of stellar population synthesis modelling very similar to those used in Gallazzi et al. (2005, 2008) and Salim et al. (2007).[3] Roughly speaking, a large library of template spectra is generated from stellar population synthesis models. To estimate the SFR of a certain observed galaxy, one would compare the galaxy's spectrum to each of the template spectra. The SFRs of the templates are then weighted based on the likelihood of the template spectra given the real spectrum, resulting in a probability distribution for the SFR. From this distribution, the final SFR of the galaxy is calculated as the expected SFR.

To estimate the sSFR when only photometric information is available, the template spectra are multiplied by the filter transmissions of the particular survey, in our case SDSS (see Fig. 1), to produce template magnitudes. These are then compared to observed ones, and the pipeline described above continues.

3.3 Description of experiments

We considered four different experimental set-ups with the common goal of estimating sSFRs and photo-z's of galaxies as accurately as possible. The first two experiments were based on the exact same galaxy sample as used for the template-based model (and can thus be compared directly), whereas the last two experiments were based on the total selected galaxy sample mentioned in Section 3.1, but with the smaller subset excluded (hereafter referred to as the *larger subset*). The experiments for sSFR and photo-z estimations were identical in set-up – only the quantity to be estimated changed.

Common to all experiments is that we used the four colours $u - g$, $g - r$, $r - i$, and $i - z$ of the galaxies, and in the experiments with feature selection we additionally included the plain magnitudes u, g, r, i, and z. The magnitudes varied from experiment to experiment, see the summary in Table 1 and detailed description further down. The data and code used for the experiments, as well as the results of these, can be found online, see Appendix B.

In each experiment, a nested CV – an inner and an outer – was used to assess the performance of the k-NN method. Both the inner and outer CV partitioned the data into 10 folds with 9 folds being used for training and the remaining fold being used for testing. For each outer CV, the 9 folds of training data were further partitioned into 10 inner folds for the inner CV. Of these 10 inner folds, 9 were used as training data and the remaining as test data in order to determine the optimal $k \in \{2, 3, 4, \ldots, 50\}$, while simultaneously doing feature selection by minimizing the root-mean-square error (RMSE). The exact number of chosen features, as well as which features were chosen, therefore varied across all folds. This simultaneous k determination and feature selection was made possible by the massively parallel graphics processing unit (GPU) implementation of the k-NN algorithm described in Gieseke et al. (2014a). Doing feature selection on the scale of this study is simply not feasible without a highly optimized k-NN implementation.

After the optimal features and optimal k were determined by the inner CV, the performance was assessed by the outer CV.

The performance of each method was therefore assessed 10 times, allowing us to calculate both the means and (population) standard deviations for each of the performance metrics discussed in Section 4. As folds in a CV procedure are not fully independent of each other, these standard deviations cannot be interpreted as strict confidence intervals.

To make the estimations by the k-NN, the template-based model (for the sSFR estimations), and the SDSS method (for the photo-z estimations) comparable, the predictions by the latter two methods were divided into the same 10 subsets as used in the outer CV of the k-NN, and the same statistics were calculated. The four experiments were devised as follows.

[3] J. Brinchmann, private communication.

Table 1. Summary of experiments. The experiments were based on the four colours $u-g$, $g-r$, $r-i$, and $i-z$, as well as the five magnitudes u, g, r, i, and z, where indicated.

Experiment	Sample size	Feature selection	Features
1	7799	No	`modelMag`; colours only
2	7799	Yes	`psfMag`, `fiberMag`, `petroMag`, `deVMag`, `expMag`, `modelMag`; colours and magnitudes
3	603 680	No	`modelMag`; colours only
4	603 680	Yes	As selected in experiment 2

Table 2. RMSEs, medians, and scatter of ΔsSFR, shown as their mean and standard deviations over the 10 CV folds for the k-NN regressions and the template-based model.

Experiment	D	RMSE/10^{-2}log(yr^{-1})	Median/10^{-2}log(yr^{-1})	Scatter, σ/10^{-2}log(yr^{-1})	η/per cent
		SDSS subset of 7799 galaxies			
1	4	29.0 ± 1.8	1.63 ± 1.20	28.9 ± 1.7	1.78 ± 0.36
2	8[a]	27.1 ± 1.5	1.52 ± 0.99	27.0 ± 1.4	1.72 ± 0.32
Template-based model		34.9 ± 1.6	− 12.4 ± 0.8	30.4 ± 1.6	3.05 ± 0.35
		SDSS subset of 603 680 galaxies			
3	4	29.6 ± 0.2	1.65 ± 0.11	29.6 ± 0.2	1.78 ± 0.04
4	8	27.4 ± 0.3	1.33 ± 0.08	27.4 ± 0.3	1.85 ± 0.05

[a]Number of features is the median of the 10 CV folds.

Experiment 1. The first experiment used the smaller subset (7799 galaxies) and used the four `modelMag` colours $u-g$, $g-r$, $r-i$, and $i-z$ as features. No feature selection was performed, but k was still optimized in each of the inner CV folds. This experiment acts as a baseline for the later feature selection.

Experiment 2. The second experiment again used the smaller subset, but this time all six types of magnitudes (`psfMag`, `fiberMag`, `petroMag`, `deVMag`, `expMag`, and `modelMag`) were used. Each type of magnitude gives rise to four colours and five magnitudes, totalling 54 features. A feature selection was performed independently for each outer CV fold to find the best feature combination.

Experiment 3. The third experiment used the larger subset. The features were again only the four `modelMag` colours, and the experiment will serve as a baseline for the k-NN performance on this larger subset.

Experiment 4. The fourth experiment also used the larger subset. The features were chosen to be the overall most informative ones found in experiment 2, based on a median ranking of the importance of each feature across the CV folds. This last experiment will test how well k-NN, with features found from a feature selection on a small data set, can extended to a much larger data set, thus assessing its performance in a 'big data' setting.

4 RESULTS AND ANALYSIS

4.1 sSFR experiments

We evaluate the sSFR experiments using the following performance metrics. We define the logarithm of the ratio of the estimated sSFR to the spectroscopically confirmed, $\Delta\text{sSFR} \equiv \log_{10}(\text{sSFR}_{\text{est}}/\text{sSFR}_{\text{spec}})$. For each CV fold m, we compute the RMSE as

$$\text{RMSE} = \sqrt{\frac{1}{|S_m|}\sum_{n\in S_m}\Delta\text{sSFR}_n^2},$$

where S_m is the test set. We also compute the median of ΔsSFR, as well as the scatter, σ, defined to be the standard deviation of ΔsSFR over the test set. Lastly, we report the fraction of catastrophic outliers, η, defined to be galaxies with $|\Delta\text{sSFR}| > 3\sigma$.

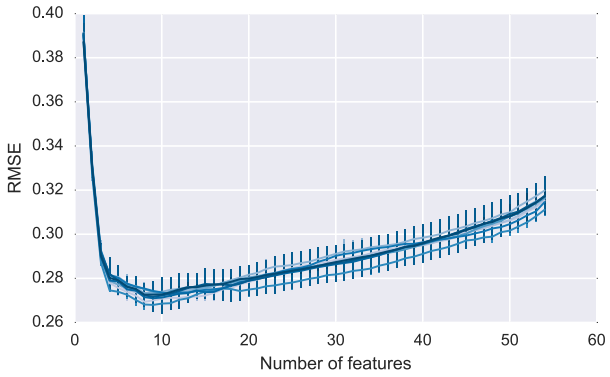


Figure 3. RMSE and one standard deviation intervals for each of the 10 CV folds of the estimated sSFRs during feature selection. A sharp decrease in error is seen as the first features are added, but it levels off quickly after the first three added features. As features continued to be added, the errors started increasing again.

Results of the sSFR experiments can be seen in Table 2. The reported values are the means and standard deviations of each performance metric over the 10 CV folds.

Comparing first the results of the experiments on the smaller subset of SDSS (experiments 1 and 2) to the result of the template-based model, we see a clear overall improvement for both experiments. In particular, the median is much improved, showing that k-NN achieves a lower bias.

In addition, doing feature selection (experiment 2) rather than simply using the four `modelMag` colours (experiment 1) further improved the estimations, though not as significantly as the improvements over the template-based model. Fig. 3 shows the RMSE and standard deviation of the sSFR estimation for each of the 10 CV folds of the smaller subset during feature selection (experiment 2). The RMSE and standard deviation are computed each time a feature is added. It is seen that by far the largest gain in accuracy happened with the addition of the first three features (which for all folds are three `modelMag` colours, see below). The error kept

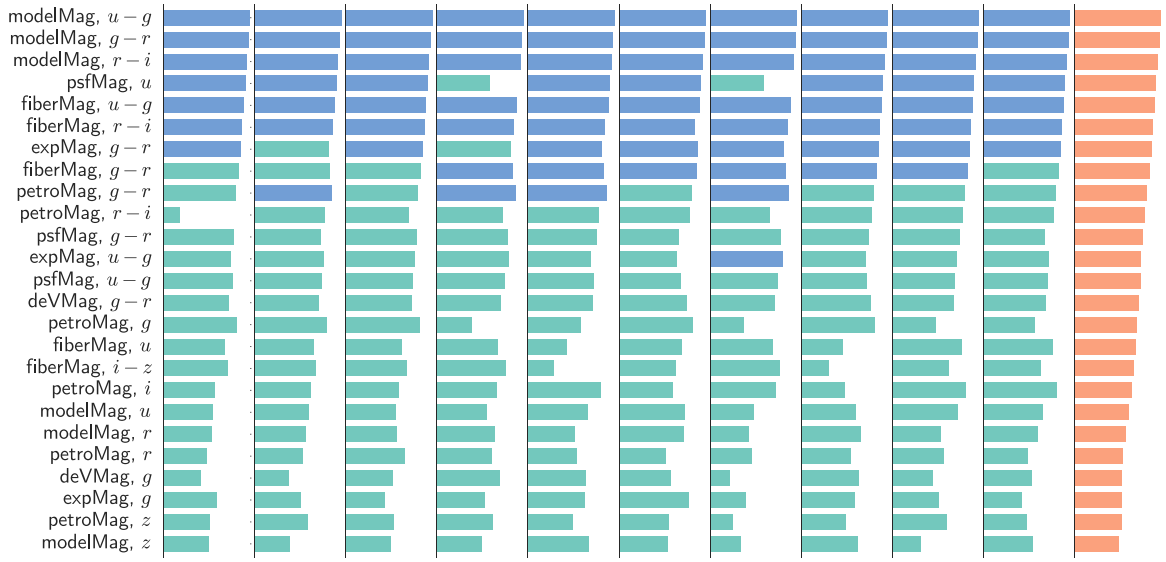


Figure 4. Ranking of the top 25 most important features from the feature selection in experiment 2. To the left are the feature names, while the rightmost column shows the median rank of each feature across all CV folds. Each of the other columns shows the feature ranking in a particular CV fold. The larger the bar for a certain feature, the more important the feature was. Blue bars show features that were chosen during the feature selection as the most informative in a particular CV fold. Because of the differences in the data used in each CV fold, the exact features selected as important, as well as the number of chosen features per fold, will vary. The number of chosen features vary from 7 through 10, with a median of 8.

decreasing until it was at its lowest at seven to ten added features after which the error started to increase. This is a very commonly seen behaviour for *k*-NN, and the reason is likely the decreasing quality of the features; as the dimensionality of the feature space increases, we are adding less informative (i.e. noisier) features. The combined effect is that the nearest neighbours to any given data point might change and the estimation will be worse as a result. It is therefore important to stop the feature selection process before the error starts increasing. The results for experiment 2 in Table 2 were achieved in exactly this way, i.e. by stopping when the RMSE was lowest.

To see which features were chosen in experiment 2, and in which order they were chosen, the results for each CV are illustrated in Fig. 4. The full list of ranked features can be seen in Fig. C1. The names of the features are shown to the left of the plot, and the middle 10 columns show the ranking of the features for each of the 10 CV folds, with a larger bar indicating that the feature was chosen earlier in the feature selection process, and thus has higher importance. A bar is coloured blue if the corresponding feature was selected in the feature selection process. Note that the amount of selected features per fold varies, as do the chosen features themselves. This is due to the differences in the data for each fold. This variation should become less prominent with an increased amount of data, as the folds will statistically become more and more similar. The rightmost column shows the median rank of each feature over all CV folds.

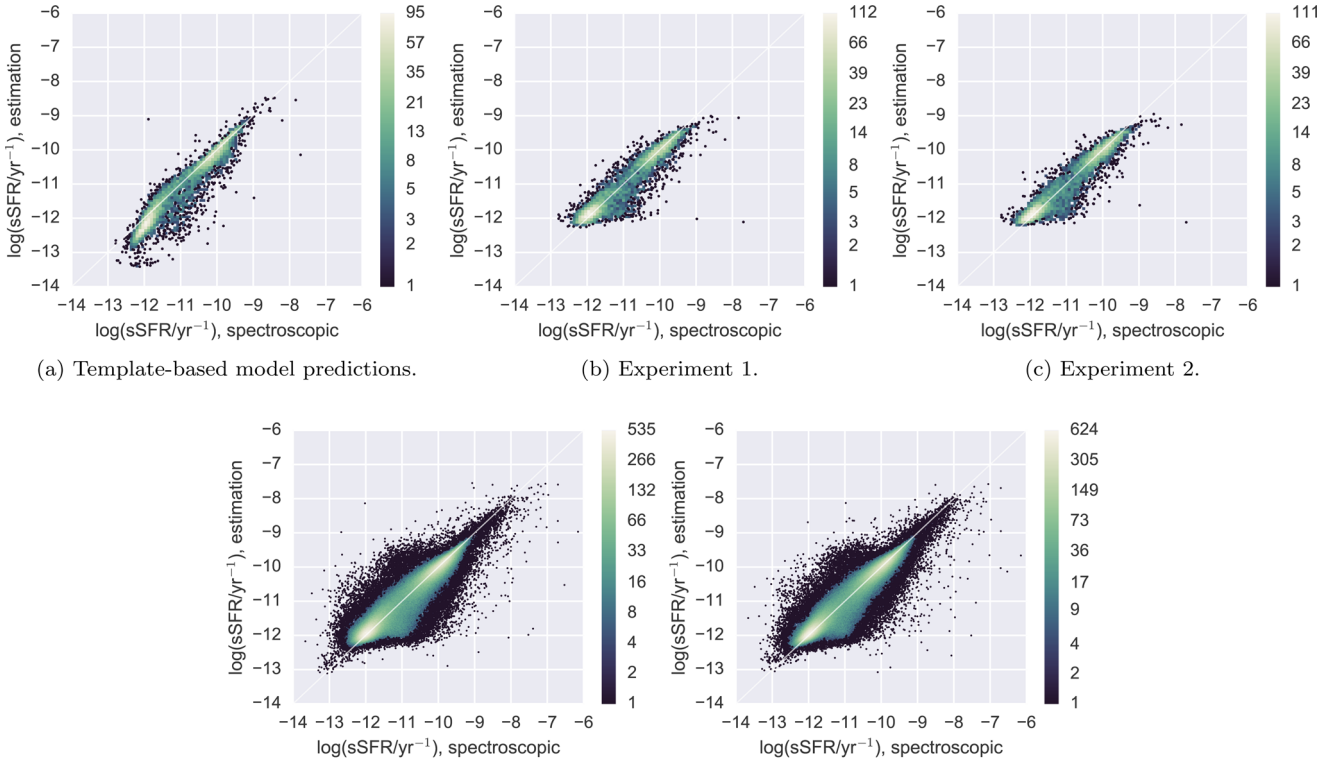
It is seen that the top six features were consistently chosen in each CV fold, except for the *u*-band `psfMag`, which seems to have been replaced by the *i* band in two folds, see Fig. C1. The `expMag` $g-r$ colour was also selected in all but two folds, though it is less clear whether it was replaced with something else. The remaining chosen features varied more, but were also consistent enough that, except for the *i*-band `psfMag` and the `expMag` $u-g$ colour, no feature below the ninth in the figure was ever chosen. For the overall most informative features (for use in experiments 2 and 4), we chose to select the top eight features from the plot, since eight is the median of the number of chosen features across the CV folds. This is just one particular choice, and one may equally well explore the benefits of using other selection criteria or ranking methods, e.g. ranking based on how often a feature was chosen. Indeed, testing various ranking and selection criteria is an obvious extension to our work, though the exact choices are unlikely to cause significant changes to the results. In summary, we chose to base both ranking and selection on the median.

Returning to the figure, it is interesting that only a single `petroMag` colour was ever selected, even though these are the magnitudes recommended by the SDSS[4] for use with low-redshift galaxies. Instead, the most prominent features were `modelMag` and `fiberMag` colours, with `modelMag` colours as the top three most informative features. This is not surprising, since the `modelMag` magnitudes are defined as either `expMag` or `deVMag` magnitudes depending on which fits the best.

Interestingly, none of the selected features use the *z* band, which can likely be explained by the band's low filter transmission, as seen in Fig. 1. This will often result in a low S/N. Also interesting is the fact that the *u* band appears in many of the most informative features, even though it also has a low S/N. The reason is likely that the band captures UV radiation from newly formed stars, thus directly measuring (part of) the SFR. This result fits well with the analysis of Davies et al. (2016), who found the NUV and *u* bands to be optimal for measuring unbiased SFRs.

Looking further down the list of selected features, we see that magnitudes and colours based on `expMag` were generally ranked much higher than their `deVMag` counterparts. This is interesting, as `modelMag`, which dominates the list of informative features, is the better fit of `expMag` and `deVMag`. This could suggest that the `modelMag` mostly resorted to a `deVMag` fit; adding `deVMag` colours (again) would not provide any new information, so the feature selection chooses to add `expMag` colours instead. Indeed,

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



(a) Template-based model predictions. (b) Experiment 1. (c) Experiment 2.

(d) Experiment 3. (e) Experiment 4.

Figure 5. Correlations between the estimated and spectroscopically determined sSFRs for the template-based model and the four experiments. The colour coding indicates the amount of galaxies in each bin.

comparing the likelihoods of the `deVMag` and `expMag` fits[5] reveals that the `deVMag` fit achieved the largest likelihood for ∼66 per cent of the galaxies in the smaller subset.

Returning to Table 2 and now considering experiment 3, which used the larger subset, but only the four `modelMag` colours, we see a performance similar to that of experiment 1, though now with significantly reduced uncertainties due to the larger sample size.

Experiment 4 also used the larger subset, but with the eight features chosen as the most informative in experiment 2 (the top eight colours and magnitudes in Fig. 4). As stated previously, the idea behind this experiment was to see how features selected on a smaller subset generalize to a larger one. This is important to know if this method is to be applied to a larger part of SDSS without any spectroscopically determined sSFRs to check for consistency with. The results from experiment 4 show that the feature selection from experiment 2 did indeed increase the performance of the method compared to using the standard colours (experiment 3). The fact that the results of experiment 4 were consistent with those of experiment 2 shows that the most informative features can indeed be determined from a smaller subset and then used on a larger. Additionally, it shows that *k*-NN regression can be an effective method for determining sSFRs from photometric data, even when the features are determined from a much smaller subset.

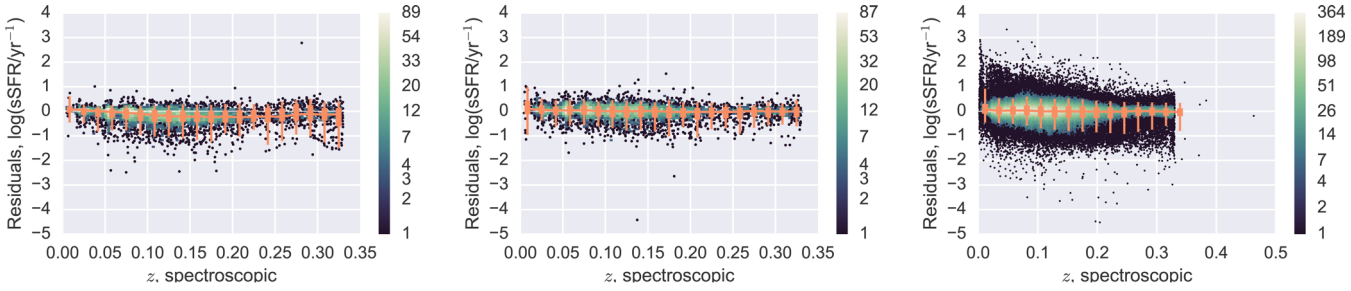
Fig. 5 shows the correlations between the spectroscopically determined sSFRs and the corresponding estimations from the template-based model as well as each of the four experiments. Looking at the estimations from the template-based model (Fig. 5a), it is immediately clear where it falls short; it seems to consistently underestimate the sSFRs of the low-sSFR galaxies. The distribution for high-sSFR galaxies also seems slightly skewed towards underestimation.

The estimations made by the *k*-NN regression (Figs 5b and c) were clearly better than those from the template-based model. The distribution for high-sSFR galaxies seems quite symmetric, while for the low-sSFR galaxies it appears slightly skewed towards overestimating the sSFRs.

The same trends can be seen in the estimations by the *k*-NN regression on the larger subset (Figs 5d and e): a symmetric mode for the high-sSFR galaxies and a slightly skewed mode for the low-sSFR galaxies, though not as pronounced as for the smaller subset.

For all *k*-NN experiments, the distribution at highest sSFRs seems to skew towards underestimation. This is likely due to the inherent inability of *k*-NN to extrapolate beyond the distribution of the training set; as there are only few data at these sSFRs, it is likely that the average of the nearest neighbours (in colour space) will drive the estimated sSFR towards lower values. Apart from choosing a different method than *k*-NN, an obvious remedy would be to include more galaxies in the training set to cover more of the colour–magnitude and sSFR space. Another possibility would be to include colours from other surveys, thereby increasing the dimensionality of the colour–magnitude space. This could potentially add the extra information needed in order to move the galaxies closer to others with similar sSFRs. Indeed, Salim et al. (2005) showed that a combination of SDSS and *GALEX* (Martin et al. 2005) photometry led to a significant improvement in the estimation of SFRs over using just

[5] Available through the `PhotoObjAll` table in the SDSS data base.



(a) Template-based model, small subset. (b) Experiment 2. (c) Experiment 4.

Figure 6. sSFR residuals as function of redshift for the two galaxy samples used in the experiments. The colour coding of the distributions indicates the amount of galaxies in each bin. The orange line shows the running median of the underlying distribution, the thick bars span the 15.87th through the 84.13th percentile ($\pm 1\sigma$), and the thin bars span the 2.28th through the 97.72th percentile ($\pm 2\sigma$). Residual plots for experiments 1 and 3 can be found in Appendix D.

SDSS photometry. It is natural to assume that this would also be the case with our method.

When looking at Fig. 5, all distributions seem to have a hump around $(-11, -12)$, where the sSFRs are somewhat underestimated. It appears as if galaxies from the green valley get mixed up with quenched galaxies. The problem also seems to be present for the template-based model, indicating that there may not be enough information in the SDSS magnitudes to distinguish these galaxies from quenched ones. Giving the galaxies a closer look would be an obvious next step to further increase the accuracy of the methods. It is, however, clear that the k-NN method works equally well for estimating sSFRs for both main-sequence and quenched galaxies, which is a rare quality for sSFR estimation methods in general.

Fig. 6 shows the sSFR residuals as function of spectroscopic redshift, with the orange line showing the running median of the underlying distribution. The thick bars span the 15.87th through the 84.13th percentile ($\pm 1\sigma$), and the thin bars span the 2.28th through the 97.72th percentile ($\pm 2\sigma$).

The template-based model (Fig. 6a) has a clear tendency to underestimate the sSFR throughout the entire redshift range. Our k-NN model (Figs 6b and c) performs a lot better, with a running median close to 0 at all redshifts. The scatter around the running median seems similar for both models, which is also apparent from Table 2.

Although the data are limited to rather low redshifts, it is reassuring to see that there appears to be no significant increase in either bias or scatter, even at the highest redshifts with our model. Note that the redshift was not part of the features used by our method. Estimation of sSFRs at all redshifts is based solely on colours and magnitudes of the galaxies.

4.1.1 Estimating uncorrected sSFRs from `fiberMag` *colours*

The colours and magnitudes are ranked based on their ability to estimate aperture-corrected sSFRs. As aperture correction is an inherently difficult task, one might expect that we would be able to obtain better accuracies by estimating uncorrected sSFRs using `fiberMag` colours and magnitudes only, since these should cover the same part of the galaxies. We briefly tried to conduct experiments 1 through 4 using only `fiberMag` colours and magnitudes for estimating uncorrected sSFRs, the results of which can be found in Appendix E, but surprisingly the results were consistently worse than for the experiments described above. All experiments performed worse than the template-based model (though these experiments are not strictly comparable, as we do not have estimations for the uncorrected sSFRs from the template-based model), and the performance also decreased significantly when going from the smaller subset to the larger. Why this is the case is not immediately obvious, but there seems to be significantly more galaxies with severely underestimated sSFRs (one to two orders of magnitude). Thus, the reason may be that the fibre is only covering the centre of many galaxies, making it more difficult to discern between ellipticals and spiral galaxies.

4.2 Redshift experiments

The accuracy of the photo-z experiments is evaluated using the following metrics. We define the normalized redshift estimation error as $\Delta z' = \Delta z/(1+z)$, where $\Delta z = z_{\rm phot} - z_{\rm spec}$. Following Ilbert et al. (2006), we define a catastrophic outlier as a galaxy with $|\Delta z'| > 0.15$ and η as being the fraction of catastrophic outliers in a given experiment. We further use the definition of the normalized median absolute deviation as $\sigma_{\rm NMAD} = 1.48 \times {\rm median}(|\Delta z'|)$. Following Dahlen et al. (2013), we define $\sigma_{\rm RMS} = \langle \Delta z'^2 \rangle^{1/2}$ and σ_O as being the $\sigma_{\rm RMS}$ after catastrophic outliers have been removed. We also evaluate the bias, given as the mean normalized error, ${\rm bias}_z = \langle \Delta z' \rangle$, once again excluding catastrophic outliers.

Table 3 presents the results obtained in the various experiments. The results are calculated by combining the results from the test sets in each of the 10 CV folds.

Considering first the experiments on the smaller subset, the SDSS method is quite consistently outperforming our experiment 1, though the differences are within or around one standard deviation. Our experiment 2, however, is consistently outperforming the SDSS method, though again the differences are mostly within one standard deviation. Comparing our experiments 1 and 2 shows a much more significant difference; the chosen features clearly outperformed the four standard colours.

Fig. 7 shows the 25 most important features obtained from the feature selection in experiment 2, with the features chosen as most informative coloured in blue. The full list of ranked features can be seen in Fig. C2. The top six features were quite consistently chosen as the most important, whereas the remaining chosen features in each CV fold are more scattered than for sSFR estimation. The number of chosen features also vary much more: from six to eleven features are chosen in the folds. The median number of selected features was seven, so the top seven features in Fig. 7 were chosen as basis for experiment 4. The varying features as well as the number of chosen features for each CV fold can be an indication that many magnitudes and colours have very similar information content. Thus, even small differences in the data sets used in each CV fold can be enough to change the features deemed most

Table 3. Results from the photo-z estimation experiments. Evaluation metrics are the bias, $\text{bias}_z = \langle \Delta z' \rangle$; the normalized root-mean-square (RMS) error, $\sigma_{\text{RMS}} = \langle \Delta z'^2 \rangle^{1/2}$; the RMS error with outliers removed, σ_O; the normalized median absolute deviation, $\sigma_{\text{NMAD}} = 1.48 \times \text{median}(|\Delta z'|)$; and the fraction of catastrophic outliers, η. The standard deviations shown are calculated over the 10 CV folds.

Experiment	D	$\text{bias}_z/10^{-4}$	$\sigma_{\text{RMS}}/10^{-2}$	$\sigma_O/10^{-2}$	$\sigma_{\text{NMAD}}/10^{-2}$	$\eta/10^{-2}$ per cent
			SDSS subset of 7799 galaxies			
1	4	-6.13 ± 9.80	2.19 ± 0.08	2.17 ± 0.07	1.72 ± 0.10	2.56 ± 5.13
2	7[a]	-1.00 ± 8.32	1.82 ± 0.09	1.81 ± 0.08	1.46 ± 0.06	2.56 ± 5.13
SDSS	4[b]	-5.84 ± 9.58	2.01 ± 0.06	1.99 ± 0.07	1.54 ± 0.06	3.85 ± 5.88
			SDSS subset of 603 680 galaxies			
3	4	3.52 ± 0.66	2.22 ± 0.03	2.20 ± 0.03	1.77 ± 0.02	2.82 ± 0.59
4	7	0.401 ± 0.524	1.72 ± 0.02	1.71 ± 0.02	1.38 ± 0.01	0.895 ± 0.372
SDSS	4[b]	5.29 ± 0.62	2.22 ± 0.02	2.12 ± 0.01	1.65 ± 0.01	8.58 ± 1.02

[a]Number of features is the median of the 10 CV folds.
[b]SDSS additionally fitted a hyperplane in order to make estimations.

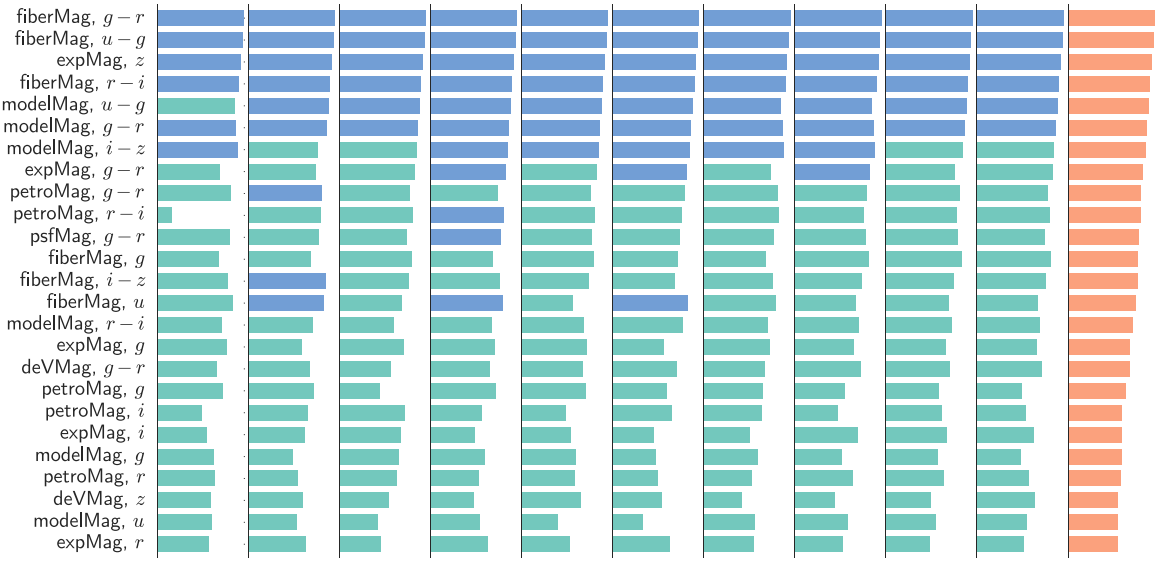


Figure 7. Ranking of the 25 most important features according the feature selection in experiment 2. To the left are the feature names, while the rightmost column shows the median rank of each feature across all CV folds. Each of the other columns shows the feature ranking in a particular CV fold. The larger the bar for a certain feature, the more important the feature was. Blue bars show features that were chosen during the feature selection as the most informative in a particular CV fold. Because of the differences in the data used in each CV fold, the exact features selected as important, as well as the number of chosen features per fold, will vary. The number of chosen features vary from 6 through 11 with a median of 7.

informative. Using a larger data set for the feature selection will likely make the chosen features more stable.

It is interesting to see that, while three of the four `modelMag` colours are among the selected features, they are not the most informative. The `fiberMag` colours appear to contain more information for photo-z estimation.

Another interesting observation is that the z-band `expMag` was chosen consistently in all CV folds. Having a single measure of the z-band magnitude therefore seems to be important for photo-z estimation. This is rather surprising, given the z band's low S/N, but shows that the observed SED is very informative in this regime.

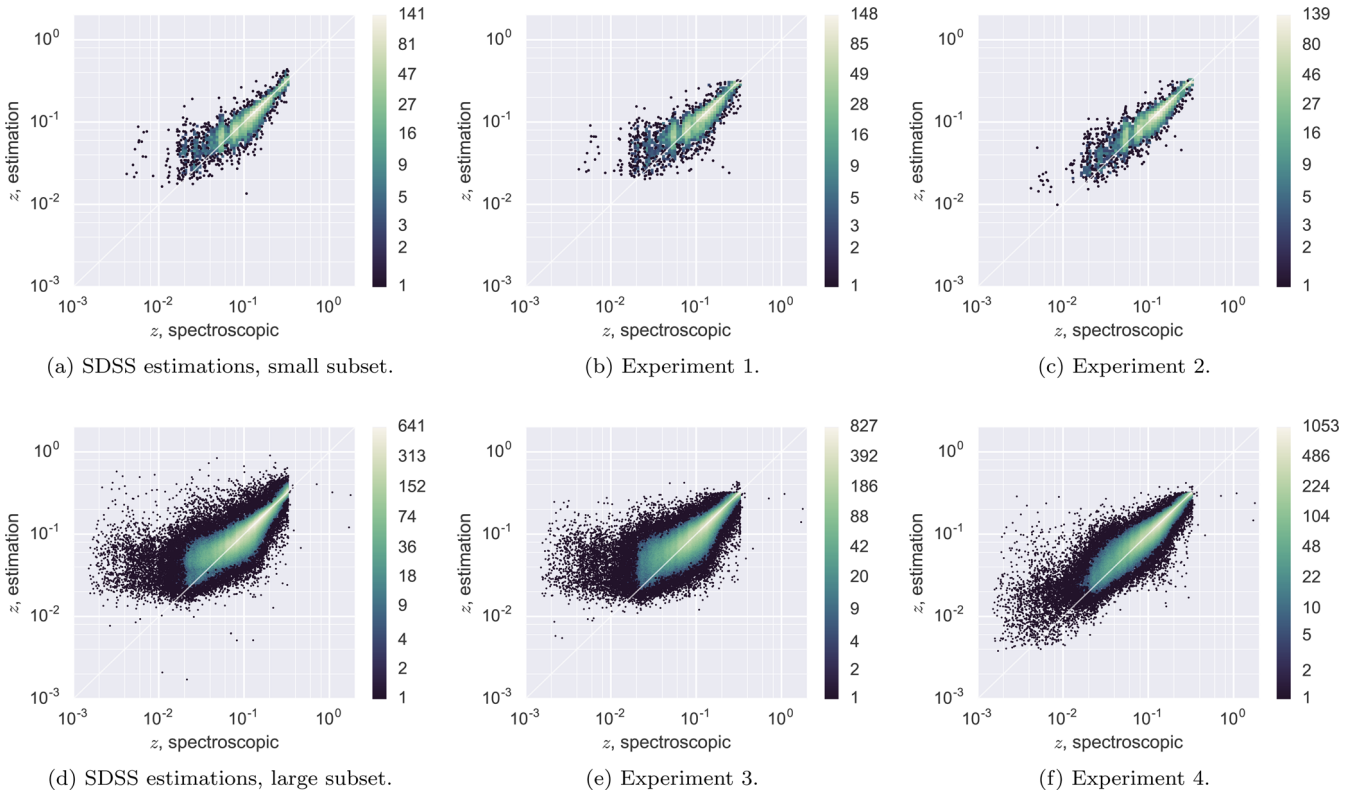
Returning again to Table 3, it is expected that the SDSS method outperform our experiment 1. Even though we use the same features, the SDSS estimate uses a hyperplane fit to the nearest 100 samples. This will act as regularization, making estimations less susceptible to outliers.

Considering now the experiments on the larger subset, the results are qualitatively as before, but with significantly reduced error bars. Overall, SDSS outperforms our experiment 3, which again uses the same features. As before, this is to be expected. Interestingly, our experiment 3 has a significantly lower outlier rate η than SDSS ($(2.82 \pm 0.59) \times 10^{-2}$ per cent versus $(8.58 \pm 1.02) \times 10^{-2}$ per cent).

Experiment 4 significantly outperformed both our experiment 3 and, more interestingly, the photo-z estimations from SDSS. We are thus able to achieve much better performance by using optimal features (found from a smaller data set) instead of the standard ones, even when using additional modelling as SDSS does.

Fig. 8 shows correlations between estimated photo-z and the spectroscopically derived redshift. The spectroscopically determined redshifts have a sharp cut around $z \sim 0.33$, after which there are only few galaxies. This is a result of our data selection.

Figs 8(a) and (d) show the photo-z estimations made by SDSS, i.e. not by our model. Figs 8(b) and (e) show the photo-z estimations done using our k-NN method, but using only the four `modelMag` colours. Finally, Figs 8(c) and (f) show the photo-z estimations done using our k-NN method, including the feature selection.



(a) SDSS estimations, small subset. (b) Experiment 1. (c) Experiment 2.

(d) SDSS estimations, large subset. (e) Experiment 3. (f) Experiment 4.

Figure 8. Correlations between the estimated photo-z and spectroscopically determined z for the SDSS photo-z method and our k-NN method. The colour coding of the distributions indicates the amount of galaxies in each bin.

Focusing first on the experiments using the smaller subset, we see that distributions resulting from our k-NN method and SDSSs are, qualitatively, quite similar. Experiment 2, which used feature selection, seems to have a slightly more symmetric distribution around the diagonal and appears to work better at the smallest redshifts, but is otherwise very similar to the other two experiments.

Turning now to the experiments using the larger subset, the SDSS method appears to result in more extreme outliers than ours. As before, this observation may be misleading. Figs 8(e) and (f) show a clear horizontal cut around $z_{\rm phot} \sim 0.33$ with only few estimated redshifts above this line. This cut is due to our data selection criteria, causing only a handful of galaxies with redshifts higher than $z \sim 0.33$ to be present in the data. As our k-NN method uses the mean of the nearest neighbours to estimate redshifts, it is restricted to always estimate redshift values within the range of the training set. Thus, a sharp cut in the redshifts of the training data means a sharp cut in the estimations, and a somewhat artificial reduction of potential outliers.

The cut is not present in the estimations from SDSS. In fact, the distribution of estimated redshifts appears unaffected above $z_{\rm phot} \sim 0.33$. One may be led to think that the method used by SDSS performs worse than ours in this region. One cannot, however, draw such a conclusion, as there are (at least) two plausible explanations for these estimations. First, the galaxies with $z_{\rm phot,SDSS} \gtrsim 0.33$ may lie in the outskirts of the data distribution in colour space, thus requiring extrapolation. If the local data distribution is not well described by a linear model, such as the one used by SDSS, the extrapolation may be of low quality. Secondly, the estimations may have been made using a data set with more high-z galaxies than ours. Indeed, it is unlikely that SDSS has employed the exact same selection criteria as we have (and only those), and as we furthermore sample randomly from our selected subset for the CVs, the two data sets will differ to some degree. If the data set used by SDSS contains more high-z galaxies, then what appears to us as extrapolations may in fact be interpolations for SDSS. Had the same galaxies been included in our data set, we might have experienced similar results.

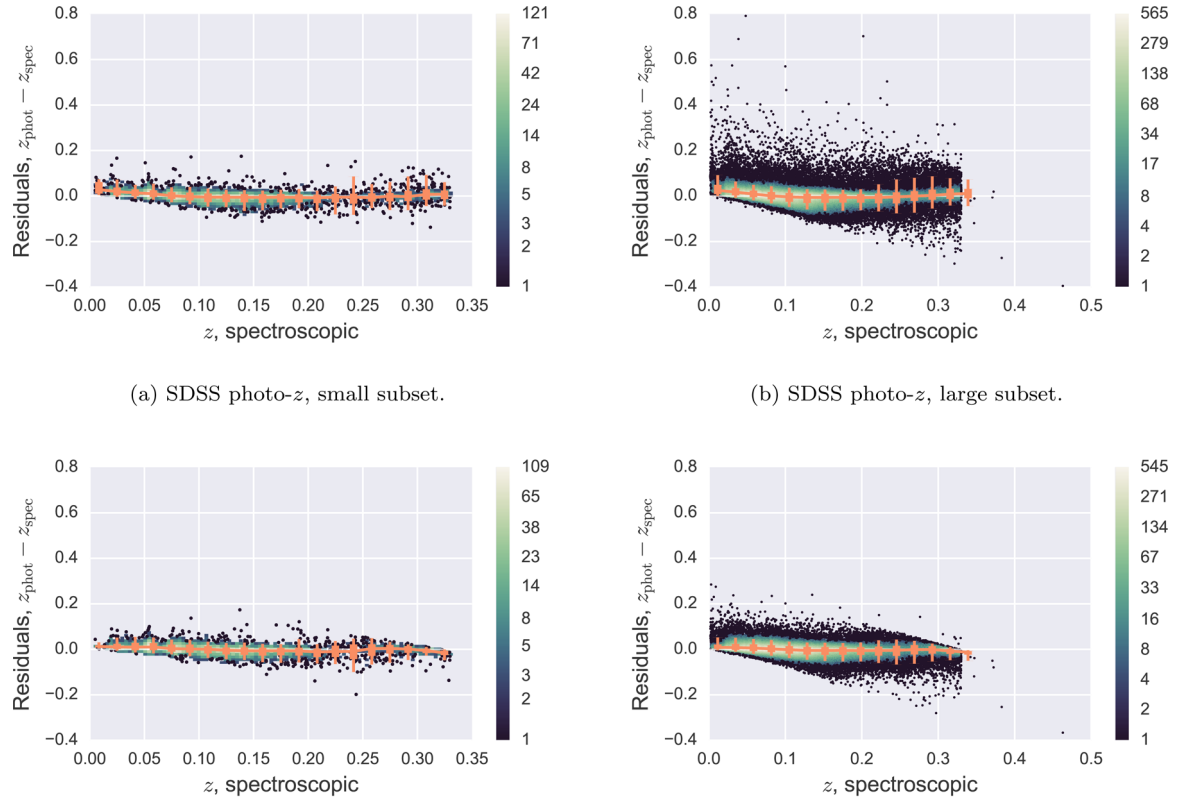
Thus, one should not draw conclusions regarding which of the methods work best based on estimations in this region.

Ignoring for a moment estimations above $z_{\rm phot} \sim 0.33$, the SDSS estimation seems to perform better than our experiment 3 (Fig. 8e), which only used the four `modelMag` colours. The SDSS photo-z distribution seems tighter around the diagonal, which is likely a result of the hyperplane fit acting as regularization. Both methods do, however, significantly overestimate at the lowest redshifts.

Fig. 8(f) shows the estimations made by our k-NN method, using the features obtained from the feature selection process in experiment 2. Compared to Fig. 8(e), there is less scatter and the distribution is significantly tighter around the diagonal. Comparing with the SDSS estimations (Fig. 8d), we perform significantly better at the lowest redshifts, with the added bonus of an overall more symmetric distribution.

Finally, Fig. 9 shows the photo-z residuals as function of spectroscopic redshift. The rather sharp slopes at $z \sim 0.3$ for our estimations are a result of the cut in the spectroscopic redshift as discussed previously. The SDSS estimations do not exhibit this slope, again suggesting that they may have been made using a data set containing more high-z galaxies, thus making extrapolation beyond $z \sim 0.3$ possible.

Considering the photo-z experiments on the small subset, there is not much difference between the estimates from SDSS (Fig. 9a) and



(a) SDSS photo-z, small subset.

(b) SDSS photo-z, large subset.

(c) Experiment 2.

(d) Experiment 4.

Figure 9. Redshift residuals as function of redshift for the two galaxy samples used in the experiments. The colour coding of the distributions indicates the amount of galaxies in each bin. The orange line shows the running median of the underlying distribution, the thick bars span the 15.87th through the 84.13th percentile ($\pm 1\sigma$), and the thin bars span the 2.28th through the 97.72th percentile ($\pm 2\sigma$). The sharp slopes seen in (c) and (d) are a consequence of the training set containing only few galaxies with $z \gtrsim 0.33$. As the k-NN method is not well suited for extrapolation, only few galaxies will have an estimated photo-z $\gtrsim$0.33. Residual plots for experiments 1 and 3 can be found in Appendix D.

those from our k-NN method (Fig. 9c). Note that the estimations from our method have been obtained using feature selection. Both estimation methods appear to overestimate the redshift at low redshifts, though it is more pronounced for the SDSS method. At higher redshifts, the both methods slightly underestimate the redshifts. At the highest redshifts, the SDSS method appears to overestimate slightly, while our k-NN method seems to underestimate the redshift. This underestimation is a consequence of the slope, as the training set used for our method contains only a few galaxies with $z \gtrsim 0.33$. Therefore, one should not conclude too much from this underestimation.

The picture is very similar when considering the experiments on the large subset (Figs 9b and d). There is a tendency to overestimate the redshift at small z and underestimate it at higher z. For both experiments, however, the median residual is always close to zero. There are a few extra galaxies at $z > 0.5$ not shown in these plots, in order to keep the main galaxy sample detailed. Both methods significantly underestimate the redshifts of these high-z galaxies with roughly the same amount.

From the plots in Fig. 9, it is clear that using just the most important features, we can achieve a similar performance to fitting a hyperplane to the nearest neighbours, though at a lower computational cost (since we do not need to locate as many nearest neighbours, and we avoid the hyperplane fitting) once the features have been determined.

5 DISCUSSION AND CONCLUSIONS

In the coming years, increasingly larger astronomical surveys will produce unprecedented amounts of data. Many of these data will require accurate estimations in near real-time, which is not feasible with traditional methods. Machine learning is well suited to address this challenge.

This work has exemplified this by showing how machine learning can be used to not only estimate sSFRs and photometric redshifts (photo-z's) of galaxies, but also to identify the most informative features for these tasks, thereby increasing accuracy further. We have shown how the simple, yet powerful non-parametric k-NN method significantly outperforms the traditional method of simulated template spectra for estimating sSFRs, achieving an RMSE of $(2.90 \pm 0.18) \times 10^{-1}$ log(yr^{-1}) (the $\pm$ values refer to the standard deviation over the non-independent CV folds) compared to a template-based method's $(3.49 \pm 0.16) \times 10^{-1}$ log(yr^{-1}), when using the exact same input features. Adding a *feature selection* to the k-NN method increased its performance, achieving an RMSE of $(2.71 \pm 0.15) \times 10^{-1}$ log(yr^{-1}). Similarly, the fraction of catastrophic outliers reduced from the template-based method's (3.05 ± 0.35) to (1.72 ± 0.32) per cent, when using k-NN and feature selection.

We see a similar pattern when considering photo-z estimation. Here, the k-NN method achieves a normalized median absolute deviation of $(1.72 \pm 0.10) \times 10^{-2}$, which reduces to

$(1.46 \pm 0.06) \times 10^{-2}$ when doing feature selection, compared to $(1.54 \pm 0.06) \times 10^{-2}$ achieved by SDSS. The method used by SDSS even included a hyperplane fit and while that improves estimations, it also significantly increases the required amount of computations per estimate.

Applying the *k*-NN method to a larger subset of SDSS of 603 680 galaxies, we achieve an RMSE of $(2.96\ |pm\ 0.02) \times 10^{-1}$ $\log(\mathrm{yr}^{-1})$ for sSFR estimation, when using the same four features as the template-based method. By using the features selected in the feature selection on the smaller subset, we are able to decrease the error further to $(2.74 \pm 0.03) \times 10^{-1}$ $\log(\mathrm{yr}^{-1})$. For photo-z estimation, we achieve a normalized median absolute deviation of $(1.77 \pm 0.02) \times 10^{-2}$, which reduces to $(1.38 \pm 0.01) \times 10^{-2}$ when doing feature selection, compared to $(1.65 \pm 0.01) \times 10^{-2}$ achieved by SDSS. This shows that not only can features selected for a smaller subset be directly transferred to a much larger one yielding similar performance, the estimations made by the selected features can even significantly outperform more computationally intensive modelling.

An advantage of a template-based method is the gain in physical knowledge from the simulations. The feature selection for the *k*-NN method can provide hints to which features contain the most information, but a deeper understanding of why these particular features contain more information requires further investigation and is outside the scope of this work. The *k*-NN method does, however, have advantages over a template-based method in that it is faster and will not be prone to errors resulting from approximations or wrong assumptions done in the model building process. This study shows that machine learning methods, here exemplified by *k*-NN regression, should be considered a viable alternative to the traditional template-based method in situations where high accuracy or computational efficiency is required. In particular, adding a feature selection step to the machine learning methods, instead of relying on traditionally used features, should be considered part of the standard toolbox.

ACKNOWLEDGEMENTS

We sincerely thank Jarle Brinchmann for providing us with photometric estimations of masses and SFRs (through private communication), and for the spectroscopic SFRs made available at http://wwwmpa.mpa-garching.mpg.de/SDSS/DR7/sfrs.html. We also thank the SDSS collaboration for making their reduced data available.

This research made use of NASA's Astrophysics Data System; NUMPY and SCIPY (Jones et al. 2001; van der Walt, Colbert & Varoquaux 2011); the IPYTHON package (Perez & Granger 2007); SCIKIT-LEARN (Pedregosa et al. 2011); PANDAS (McKinney 2010); MATPLOTLIB, a PYTHON library for publication quality graphics (Hunter 2007); and SEABORN (Waskom et al. 2016).

KSS, CI, and KSP gratefully acknowledge support from The Danish Council for Independent Research | Natural Sciences through the project 'Surveying the sky using machine learning'.

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the US Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS website is http://www.sdss.org/.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The participating institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

REFERENCES

Abazajian K. N. et al., 2009, ApJS, 182, 543
Abdalla F. B., Banerji M., Lahav O., Rashkov V., 2011, MNRAS, 417, 1891
Almosallam I. A., Lindsay S. N., Jarvis M. J., Roberts S. J., 2016, MNRAS, 455, 2387
Altman N. S., 1992, Am. Stat., 46, 175
Arnouts S. et al., 2013, A&A, 558, A67
Balogh M. L. et al., 2016, MNRAS, 456, 4364
Benítez N., 2000, ApJ, 536, 571
Bentley J. L., 1975, Commun. ACM, 18, 509
Bolzonella M., Miralles J.-M., Pelló R., 2000, A&A, 363, 476
Brammer G. B., van Dokkum P. G., Coppi P., 2008, ApJ, 686, 1503
Brinchmann J., Charlot S., White S. D. M., Tremonti C., Kauffmann G., Heckman T., Brinkmann J., 2004, MNRAS, 351, 1151
Budavári T., Szalay A. S., Connolly A. J., Csabai I., Dickinson M., 2000, AJ, 120, 1588
Carnero A., Sánchez E., Crocce M., Cabré A., Gaztañaga E., 2012, MNRAS, 419, 1689
Carrasco Kind M., Brunner R. J., 2013, MNRAS, 432, 1483
Cayton L., 2012, Accelerating Nearest Neighbor Search on Manycore Systems. IEEE, Piscataway, NJ, p. 402
Charlot S., Kauffmann G., Longhetti M., Tresse L., White S. D. M., Maddox S. J., Fall S. M., 2002, MNRAS, 330, 876
Chen Y.-M., Wild V., Kauffmann G., Blaizot J., Davis M., Noeske K., Wang J.-M., Willmer C., 2009, MNRAS, 393, 406
Collister A. A., Lahav O., 2004, PASP, 116, 345
Conroy C., 2013, ARA&A, 51, 393
Csabai I., Dobos L., Trencséni M., Herczegh G., Józsa P., Purger N., Budavári T., Szalay A. S., 2007, Astron. Nachr., 328, 852
D'Isanto A., Cavuoti S., Brescia M., Donalek C., Longo G., Riccio G., Djorgovski S. G., 2016, MNRAS, 457, 3119
da Cunha E., Charlot S., Elbaz D., 2008, MNRAS, 388, 1595
Dahlen T. et al., 2013, ApJ, 775, 93
Davies L. J. M. et al., 2016, MNRAS, 461, 458
Everson R. M., Fieldsend J. E., 2004, A Variable Metric Probabilistic *k*-Nearest-Neighbours Classifier. Springer-Verlag, Berlin, p. 654. Available at: http://dx.doi.org/10.1007/978-3-540-28651-6_96
Fillingham S. P., Cooper M. C., Wheeler C., Garrison-Kimmel S., Boylan-Kolchin M., Bullock J. S., 2015, MNRAS, 454, 2039
Fukugita M., Ichikawa T., Gunn J. E., Doi M., Shimasaku K., Schneider D. P., 1996, AJ, 111, 1748
Gallazzi A., Charlot S., Brinchmann J., White S. D. M., Tremonti C. A., 2005, MNRAS, 362, 41
Gallazzi A., Brinchmann J., Charlot S., White S. D. M., 2008, MNRAS, 383, 1439
Garcia V., Debreuve E., Nielsen F., Barlaud M., 2010, in Proc. IEEE Int. Conf. Image Process, IEEE, Piscataway, NJ, p. 3757
Geach J. E., 2012, MNRAS, 419, 2633
Gieseke F., Posterer K. L., Oancea C., Igel C., 2014a, in Wermter S., Weber C., Duch W., Honkela T., Koprinkova-Hristova P., Magg S., Palm G., Villa A. E. P., eds, Proc. Eur. Symp. Artificial Neural Networks, Speedy

Greedy Feature Selection: Better Redshift Estimation via Massive Parallelism. Comput. Intell. Mach. Learn., p. 87
Gieseke F., Heinermann J., Oancea C., Igel C., 2014b, in Xing E. P., Jebara T., eds, Proc. Int. Conf. Mach. Learn, 31st International Conference on Machine Learning, Journal of Machine Learning Research, p. 172
Graham M. J., Djorgovski S. G., Mahabal A. A., Donalek C., Drake A. J., 2013, MNRAS, 431, 2371
Guyon I., Elisseeff A., 2003, J. Mach. Learn. Res., 3, 1157
Hastie T., Tibshirani R., Friedman J., 2009, The Elements of Statistical Learning, 2nd edn. Springer, New York
Hildebrandt H. et al., 2010, A&A, 523, A31
Ho S. et al., 2012, ApJ, 761, 14
Holmes C. C., Adams N. M., 2002, J. R. Stat. Soc. B: Stat. Methodol., 64, 295
Hoyle B., Rau M. M., Zitlau R., Seitz S., Weller J., 2015, MNRAS, 449, 1275
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Ilbert O. et al., 2006, A&A, 457, 841
Indyk P., Motwani R., 1998, in Proc. 30th Annu. ACM Symp. Theory of Computing. ACM, New York, p. 604
James G., Witten D., Hastie T., Tibshirani R., 2013, An Introduction to Statistical Learning. Springer Texts in Statistics Vol. 103, Springer, New York
Jones E. et al., 2001, SciPy: Open Source Scientific Tools for Python. Available at: http://www. scipy. org/
Kauffmann G. et al., 2003, MNRAS, 341, 33
Kennicutt R. C., Jr, 1998, ARA&A, 36, 189
Kennicutt R. C., Evans N. J., 2012, ARA&A, 50, 531
Kremer J., Gieseke F., Pedersen K. S., Igel C., 2015, Astron. Comput., 12, 67
Kügler S. D., Polsterer K., Hoecker M., 2015, A&A, 576, A132
Li L., Zhang Y., Zhao Y., 2008, Sci. China Ser. G: Phys. Mech. Aston., 51, 916
McKinney W., 2010, in van der Walt S., Millman J., eds, Proc. 9th Python in Sci. Conf., p. 51
Manocha S., Girolami M., 2007, Pattern Recognit. Lett., 28, 1818
Maraston C., Pforr J., Renzini A., Daddi E., Dickinson M., Cimatti A., Tonini C., 2010, MNRAS, 407, 830
Martin D. C. et al., 2005, ApJ, 619, L1
Nakasato N., 2012, J. Comput. Sci., 3, 132
Noll S., Burgarella D., Giovannoli E., Buat V., Marcillac D., Muñoz-Mateos J. C., 2009, A&A, 507, 1793
Pacifici C. et al., 2015, MNRAS, 447, 786
Padmanabhan N. et al., 2007, MNRAS, 378, 852
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Perez F., Granger B. E., 2007, Comput. Sci. Eng., 9, 21
Polsterer K. L., Zinn P.-C., Gieseke F., 2013, MNRAS, 428, 226
Polsterer K. L., Gieseke F., Igel C., Goto T., 2014, in Manset N., Forshay P., eds, ASP Conf. Ser. Vol. 485, Astronomical Data Analysis Software and Systems XXIII. Astron. Soc. Pac., San Francisco, p. 425
Richards J. W., Freeman P. E., Lee A. B., Schafer C. M., 2009, MNRAS, 399, 1044
Rimoldini L. et al., 2012, MNRAS, 427, 2917
Salim S. et al., 2005, ApJ, 619, L39
Salim S. et al., 2007, ApJS, 173, 267
Sánchez C. et al., 2014, MNRAS, 445, 1482
Smith D. J. B., Hayward C. C., 2015, MNRAS, 453, 1597
Solorio T., Fuentes O., Terlevich R., Terlevich E., 2005, MNRAS, 363, 543
Stensbo-Smidt K., Igel C., Zirm A., Pedersen K. S., 2013, in Hu X. et al., eds, Proc. IEEE Int. Conf. Big Data, 2013 IEEE International Conference on Big Data. IEEE, Piscataway, NJ, p. 141
Tsalmantza P. et al., 2007, A&A, 470, 761
van der Walt S., Colbert S. C., Varoquaux G., 2011, Comput. Sci. Eng., 13, 22
Walcher J., Groves B., Budavári T., Dale D., 2011, Ap&SS, 331, 1
Waskom M. et al., 2016, seaborn: v0.7.0. Available at: http://doi.org/10.5281/zenodo.54844
Weinberger K. Q., Saul L. K., 2009, J. Mach. Learn. Res., 10, 207
Wetzel A. R., Tinker J. L., Conroy C., 2012, MNRAS, 424, 232
Wetzel A. R., Tinker J. L., Conroy C., van den Bosch F. C., 2013, MNRAS, 432, 336
Wetzel A. R., Tollerud E. J., Weisz D. R., 2015, ApJ, 808, L27
Wheeler C., Phillips J. I., Cooper M. C., Boylan-Kolchin M., Bullock J. S., 2014, MNRAS, 442, 1396
Williams R. J., Quadri R. F., Franx M., van Dokkum P., Labbé I., 2009, ApJ, 691, 1879
Wuyts S. et al., 2011, ApJ, 738, 106
Wuyts S. et al., 2013, ApJ, 779, 135
Xu X., Ho S., Trac H., Schneider J., Poczos B., Ntampaka M., 2013, ApJ, 772, 147
York D. G. et al., 2000, AJ, 120, 1579

APPENDIX A: MASSIVELY PARALLEL GREEDY FEATURE SELECTION

While greedy procedures such as forward or backward feature selection are significantly faster than the exhaustive search for the best-performing features, they can still be very time-consuming, even on training sets of moderate sizes. One way to accelerate such a feature selection step is to speedup the involved nearest neighbour computations. In the literature, various techniques can be found for this task. Typical methods are *k-d trees* (Bentley 1975) or *locality-sensitive hashing* (Indyk & Motwani 1998). However, such tools either perform poorly in higher dimensions or only yield approximate answers. A recent trend in data analytics is to resort to (exact) parallel implementations for many-core devices such as today's GPUs. For instance, Garcia et al. (2010) make use of highly tuned GPUmatrix multiplication libraries for nearest neighbour search. Other schemes are based on, e.g. adapted spatial search structures (Cayton 2012; Nakasato 2012; Gieseke et al. 2014b).

For the work at hand, we make use of a massively parallel matrix-based implementation that addresses incremental feature selection and nearest neighbour models recently proposed by Gieseke et al. (2014a). For the sake of completeness, we briefly outline the general workflow of the implementation: The general workflow for the case of forward selection is sketched in Algorithm 1. For a given training set S of labelled samples, start with an empty distance matrix $\mathbf{M} \in \mathbb{R}^{N \times N}$ that contains the current distances between all pairs of training samples. Further, the array `selected_dimensions` indicating the selected features and the array `val_errors` are initialized. The forward feature selection process starts in Step 4: The procedure GetValidationErrors computes, for each dimension j that has not yet been selected (i.e. `selected_dimensions`[j]=0), the CV error for the case of dimension j being 'added' to the current set of features. These values are stored in the array `val_errors` and the procedure GetMinDim returns the index of the smallest error contained in it (thus, $i_{\min}$ corresponds to the dimension whose addition leads to the smallest CV error). Afterwards, both `selected_dimensions` and $\mathbf{M}$ are updated accordingly, where $\mathbf{M}^{i_{\min}}$ denotes the all-pairs distance matrix based on dimension $i_{\min}$ only.

The procedure GetValidationErrors returns the validation errors for all dimensions that have not yet been selected and contributes most to the overall runtime. For each such dimension j, it computes a matrix $\widehat{\mathbf{M}} = \mathbf{M} + \mathbf{M}^j$ containing all pairwise distances with the distances of dimension j being 'added on the fly' to the distances that correspond to the previously selected dimensions. This intermediate training set is then used to compute the CV error for the currently selected set of dimensions. It turns out that this procedure and the overall workflow is particularly well suited for a massively

Algorithm 1 FORWARDSELECTION($S, \bar{d}$)

Require: Training set $S = \{(x_1, y_1), \ldots, (x_N, y_N)\} \subset \mathbb{R}^D \times \mathbb{R}$ and a number $\bar{d} < D$ of desired features.
Ensure: Array `selected_dimensions` with selected features.
1: Initialize empty distance matrix $\mathbf{M} \in \mathbb{R}^{N \times N}$;
2: `int selected_dimensions[D] = {0, ..., 0};`
3: `float val_errors[D];`
4: **for** $i = 1, \ldots, \bar{d}$ **do**
5: `val_errors` = GETVALIDATIONERRORS($\mathbf{M}$);
6: $i_{\min}$ = GETMINDIM(`val_errors`);
7: `selected_dimensions`[$i_{\min}$] = 1;
8: $\mathbf{M} = \mathbf{M} + \mathbf{M}^{i_{\min}}$;
9: **end for**
10: **return** `selected_dimensions`

parallel implementation. Basically, one can parallelize the search over all dimensions that have not yet been selected as well as over the computations of the induced CV errors. By using a standard GPU device, one can reduce the runtime by a factor of up to 150 compared to single-core CPU implementation, hence, reducing the practical runtime needed from hours to minutes only. We refer to Gieseke et al. (2014a) for the technical details and an experimental analysis of the runtimes for typical astronomical data sets.

APPENDIX B: OBTAINING CODE AND DATA

We want to make the results presented in this paper as reproducible as possible, so we are releasing the code, the data obtained from SDSS, and the results of the experiments. The code for the GPU implementation of the nearest neighbours search is available at GitHub: https://github.com/gieseke/speedynn. The scripts and data for reproducing the main results of this paper can be found at http://image.diku.dk/kstensbo/papers/1606.01/. The page contains a step-by-step guide to setting up the software and recreating the main results presented in this paper.

APPENDIX C: RESULTS FROM FEATURE SELECTION

Fig. C1 shows the full feature ranking for the sSFR estimation done in experiment 2.

Fig. C2 shows the full feature ranking for the photo-z estimation done in experiment 2.

APPENDIX D: RESIDUAL PLOTS

Residual plots for sSFR experiments 1 and 3 can be seen in Fig. D1 together with residuals of the template-based model, for comparison.

Residual plots for photo-z's experiments 1 and 3 can be seen in Fig. D2 together with residuals of the SDSS method for the same data sets, for comparison.

APPENDIX E: ESTIMATIONS FROM `fiberMag` EXPERIMENTS

Estimations from our experiments using `fiberMag` colours and magnitudes to estimate uncorrected sSFRs can be seen in Fig. E1. Experiments 1 and 3 used only the four `fiberMag` colours ($u - g$, $g - r$, $r - i$, and $i - z$) as features. The smaller subset of 7799 galaxies was used in experiment 1, whereas the larger subset of 603 680 galaxies was used in experiment 3. Experiment 2 used feature selection to choose the most informative features among the five `fiberMag` magnitudes and the four `fiberMag` colours. The experiment was done on the smaller subset. The only features selected were the $u - g$, $g - r$, and $r - i$ colours, and these were consistently selected in all CV folds. Experiment 4 used the features found in experiment 2, but now applying them to the larger subset.

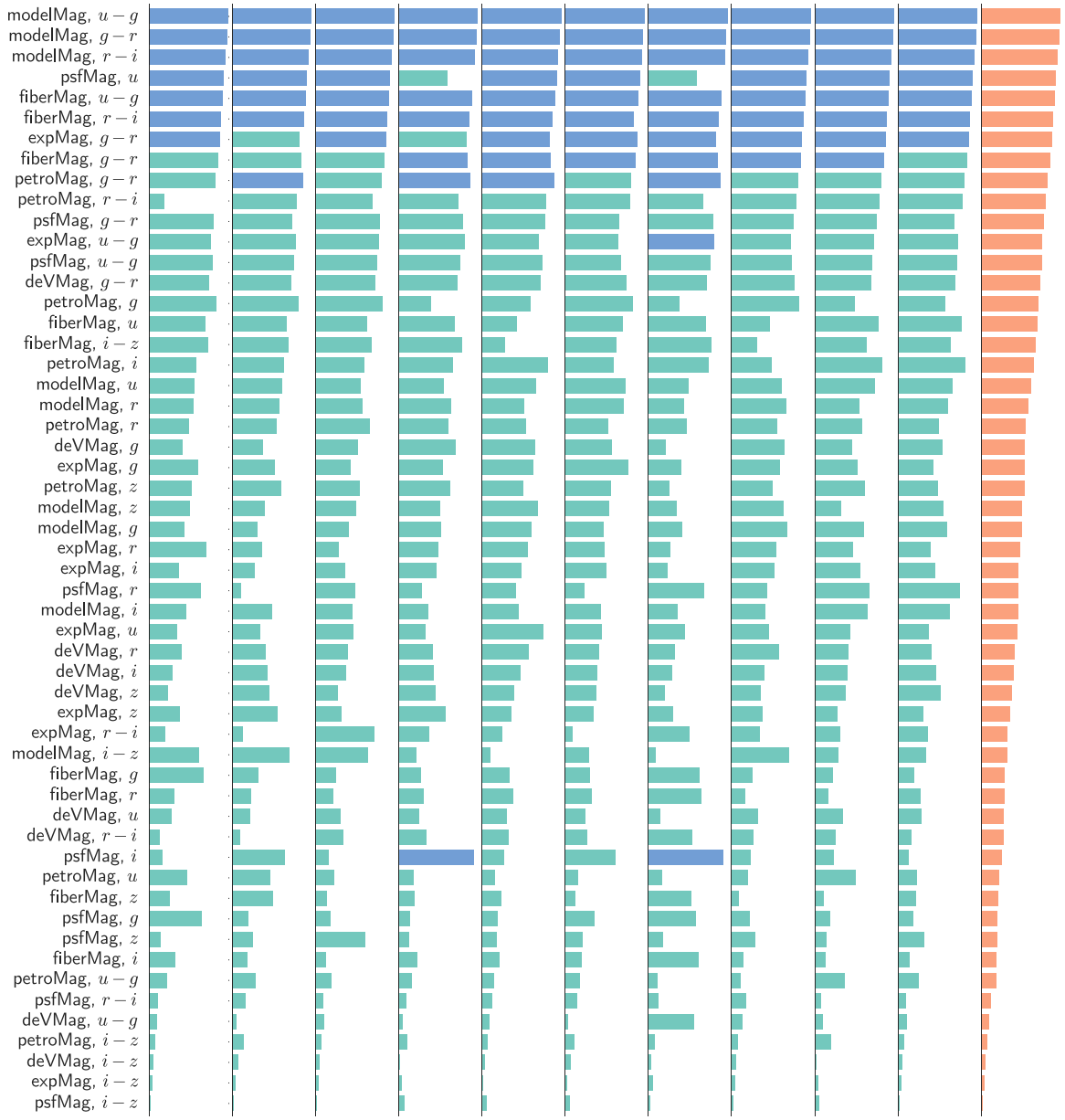


Figure C1. Ranking of features for sSFR estimation according the feature selection in experiment 2. To the left are the feature names, while the rightmost column shows the median rank of each feature across all CV folds. Each of the other columns shows the feature ranking in a particular CV fold. The larger the bar for a certain feature, the more important the feature was. Blue bars show features that were picked out during the feature selection as the most informative in a particular CV fold. Because of the differences in the data used in each CV fold, the exact features picked out as important, as well as the number of chosen features per fold, will vary. The number of chosen features vary between 7 and 10 with a median of 8.

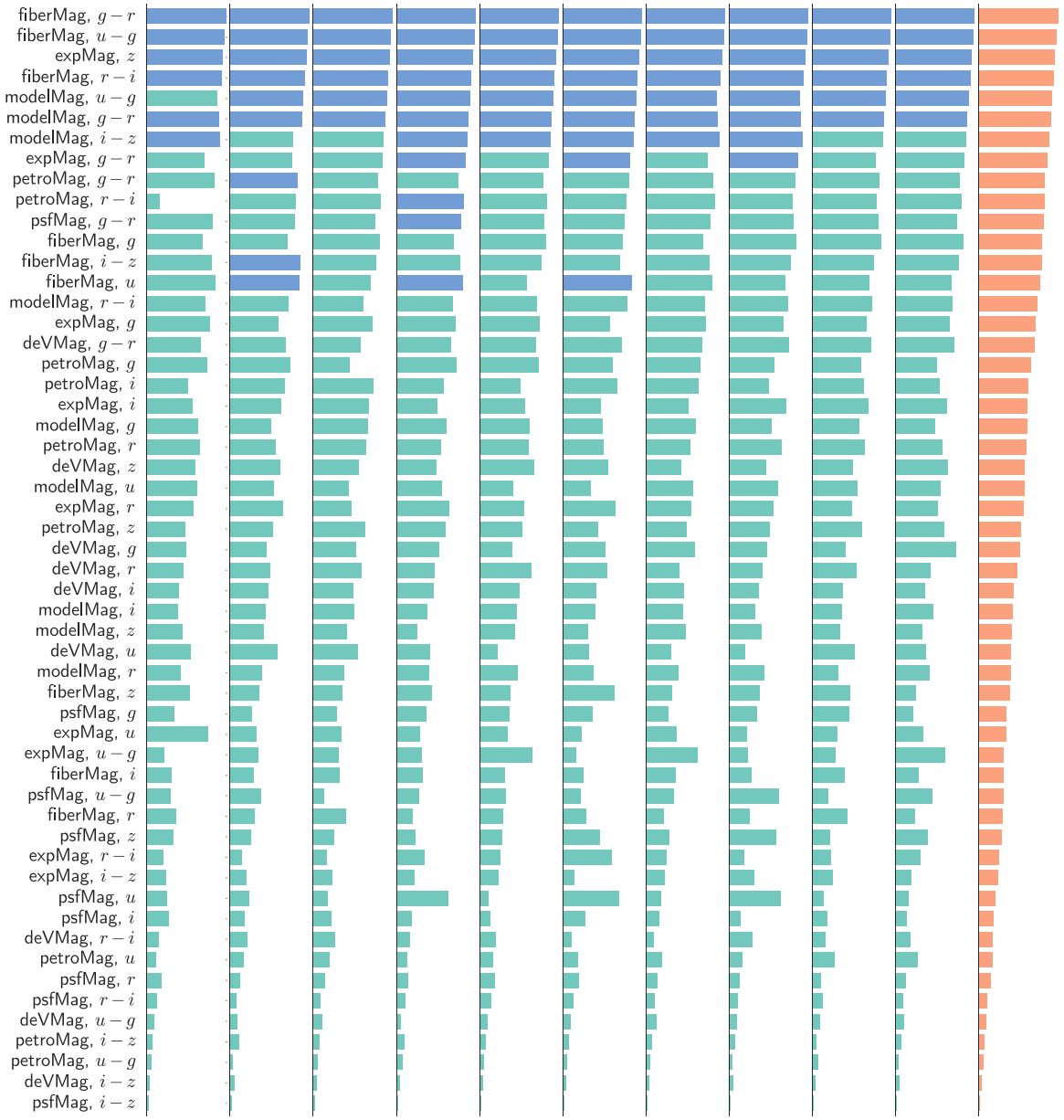
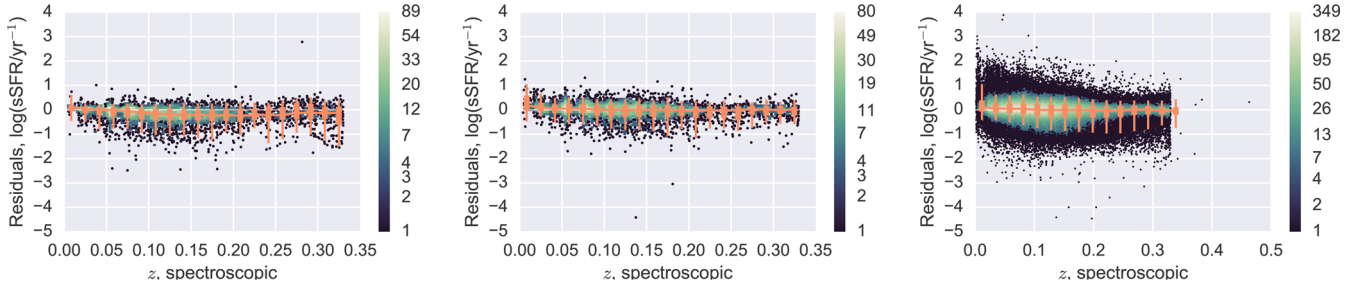
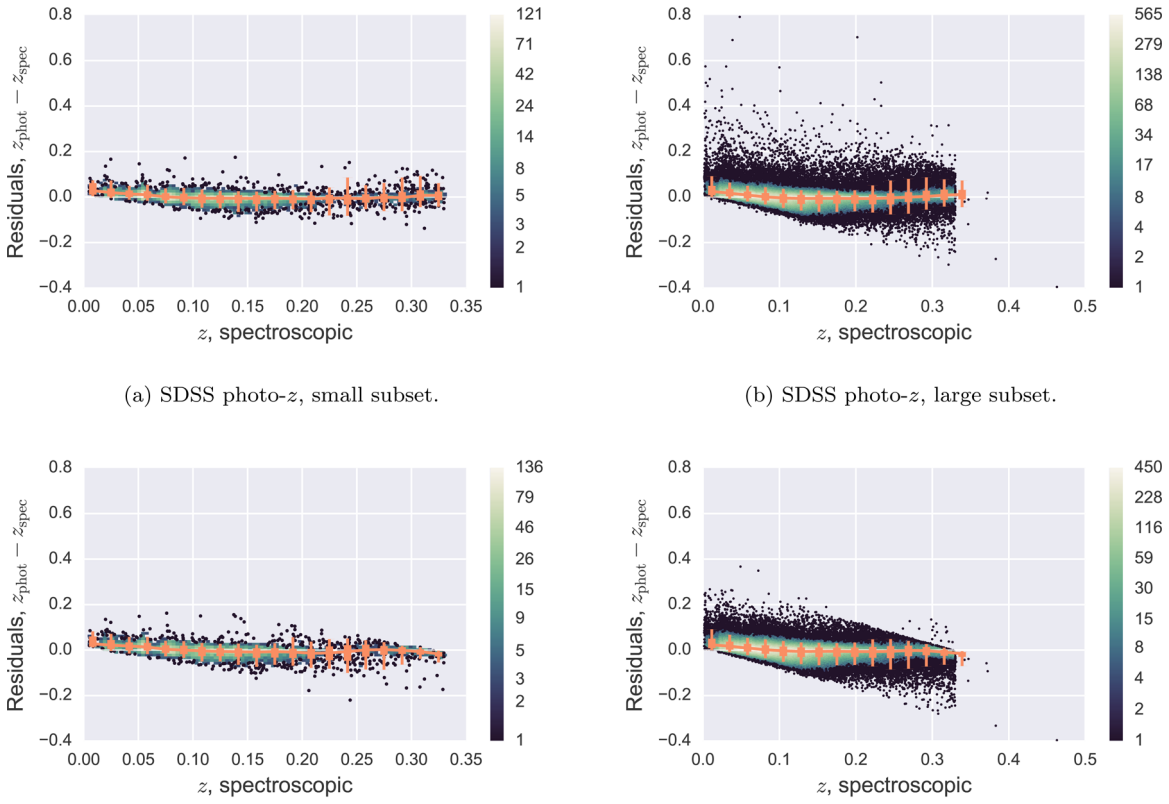


Figure C2. Ranking of features for photo-z estimation according the feature selection in experiment 2. To the left are the feature names, while the rightmost column shows the median rank of each feature across all CV folds. Each of the other columns shows the feature ranking in a particular CV fold. The larger the bar for a certain feature, the more important the feature was. Blue bars show features that were picked out during the feature selection as the most informative in a particular CV fold. Because of the differences in the data used in each CV fold, the exact features picked out as important, as well as the number of chosen features per fold, will vary. The number of chosen features vary between 6 and 11 with a median of 7.



(a) Template-based model, small subset. (b) Experiment 1. (c) Experiment 3.

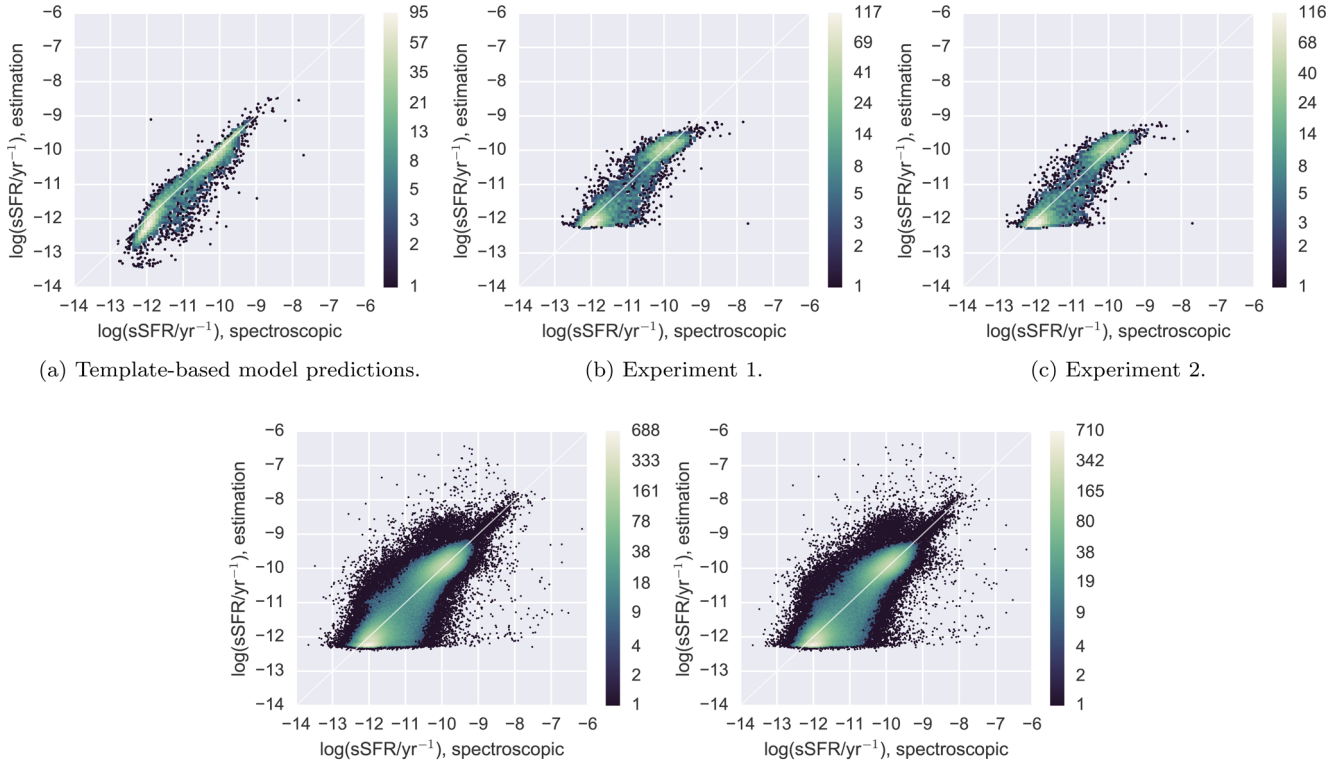
Figure D1. sSFR residuals as function of redshift for the two galaxy samples used in the experiments. The colour coding of the distributions indicates the amount of galaxies in each bin. The orange line shows the running median of the underlying distribution, the thick bars span the 15.87th through the 84.13th percentile ($\pm 1\sigma$), and the thin bars span the 2.28th through the 97.72th percentile ($\pm 2\sigma$).



(a) SDSS photo-z, small subset. (b) SDSS photo-z, large subset.

(c) Experiment 1. (d) Experiment 3.

Figure D2. Redshift residuals as function of redshift for the two galaxy samples used in the experiments. The colour coding of the distributions indicates the amount of galaxies in each bin. The orange line shows the running median of the underlying distribution, the thick bars span the 15.87th through the 84.13th percentile ($\pm 1\sigma$), and the thin bars span the 2.28th through the 97.72th percentile ($\pm 2\sigma$). The sharp slopes seen in (c) and (d) are a consequence of the training set containing only few galaxies with $z \gtrsim 0.33$. As the k-NN method is not well suited for extrapolation, only few galaxies will have an estimated photo-z $\gtrsim$0.33.



(a) Template-based model predictions. (b) Experiment 1. (c) Experiment 2.

(d) Experiment 3. (e) Experiment 4.

Figure E1. Correlations between the estimated and spectroscopically determined sSFRs for the template-based model (using aperture-corrected sSFRs) and the four experiments (using uncorrected sSFRs).

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

Bibliography

Abraham, R. G., N. R. Tanvir, B. X. Santiago, R. S. Ellis, K. Glazebrook, and S. van den Bergh (1996). Galaxy morphology to I=25 mag in the Hubble Deep Field. *Monthly Notices of the Royal Astronomical Society 279*, L47–L52.

Abraham, R. G., F. Valdes, H. K. C. Yee, and S. van den Bergh (1994). The morphologies of distant galaxies. 1: an automated classification system. *Astrophysical Journal 432*, 75–90.

Abraham, R. G., S. van den Bergh, and P. Nair (2003). A New Approach to Galaxy Morphology. I. Analysis of the Sloan Digital Sky Survey Early Data Release. *Astrophysical Journal 588*, 218–229.

Bershady, M. A., A. Jangren, and C. J. Conselice (2000). Structural and Photometric Classification of Galaxies. I. Calibration Based on a Nearby Galaxy Sample. *Astronomical Journal 119*, 2645–2663.

Bertin, E. and S. Arnouts (1996). SExtractor: Software for source extraction. *Astronomy and Astrophysics Supplement 117*, 393–404.

Bishop, C. M. (2009). *Pattern Recognition and Machine Learning*. Springer.

Calcino, J. and T. Davis (2016). The need for accurate redshifts in supernova cosmology. arXiv:1610.07695.

Conselice, C. J. (2003). The Relationship between Stellar Light Distributions of Galaxies and Their Formation Histories. *Astrophysical Journal Supplement 147*, 1–28.

Dalal, N. and B. Triggs (2005). Histograms of Oriented Gradients for Human Detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Volume 1, pp. 886–893. IEEE.

Fukugita, M., T. Ichikawa, J. E. Gunn, M. Doi, K. Shimasaku, and D. P. Schneider (1996). The Sloan Digital Sky Survey Photometric System. *Astronomical Journal 111*, 1748.

Griffin, L. D. and M. Lillholm (2007). Feature category systems for 2nd order local image structure induced by natural image statistics and otherwise. In *Human Vision and Electronic Imaging*.

Guyon, I. and A. Elisseeff (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research 3*, 1157–1182.

Hubble, E. (1929). A relation between distance and radial velocity among Extra-Galactic nebulae. *Proceedings of the National Academy of Sciences of the United States of America 17*, 168–173.

Hubble, E. P. (1926). Extragalactic nebulae. *The Astrophysical Journal 64*, 321.

Kauffmann, G., T. M. Heckman, S. D. M. White, S. Charlot, C. Tremonti, E. W. Peng, M. Seibert, J. Brinkmann, R. C. Nichol, M. SubbaRao, and D. York (2003). The dependence of star formation history and internal structure on stellar mass for 10 5 low-redshift galaxies. *Monthly Notices of the Royal Astronomical Society 341*(1), 54–69.

Kingma, D. P., D. J. Rezende, S. Mohamed, and M. Welling (2014). Semi-supervised learning with deep generative models. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, Cambridge, MA, USA, pp. 3581–3589. MIT Press.

Koenderink, J. and A. V. Doorn (1999). The structure of locally orderless images. *International Journal of Computer Vision 31*, 159–168.

Koenderink, J. J. and A. J. van Doorn (1992). Surface shape and curvature scales. *Image and Vision Computing 10*(8), 557–564.

Kremer, J., K. Stensbo-Smidt, F. Gieseke, K. Steenstrup Pedersen, and C. Igel (2016). Big universe, big data: machine learning and image analysis for astronomy. *IEEE Intelligent Systems*. Accepted for publication, September 2016.

Lawrence, A., S. J. Warren, O. Almaini, A. C. Edge, N. C. Hambly, R. F. Jameson, P. Lucas, M. Casali, A. Adamson, S. Dye, J. P. Emerson, S. Foucaud, P. Hewett, P. Hirst, S. T. Hodgkin, M. J. Irwin, N. Lodieu, R. G. McMahon, C. Simpson, I. Smail, D. Mortlock, and M. Folger (2007). The UKIRT Infrared Deep Sky Survey (UKIDSS). *Monthly Notices of the Royal Astronomical Society 379*, 1599–1617.

Li, J., K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu (2016). Feature selection: A data perspective. arXiv:1601.07996.

Lindeberg, T. (1994). *Scale-Space Theory in Computer Vision*. Norwell, MA, USA: Kluwer Academic Publishers.

Lindeberg, T. (1998). Feature detection with automatic scale selection. *International Journal of Computer Vision 30*(2), 79–116.

Lintott, C. J., K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg (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.

Lotz, J. M., J. Primack, and P. Madau (2004). A New Nonparametric Approach to Galaxy Morphological Classification. *Astronomical Journal 128*, 163–182.

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision 60*(2), 91–110.

Mortlock, D. J., S. J. Warren, B. P. Venemans, M. Patel, P. C. Hewett, R. G. McMahon, C. Simpson, T. Theuns, E. A. Gonzáles-Solares, A. Adamson, S. Dye, N. C. Hambly, P. Hirst, M. J. Irwin, E. Kuiper, A. Lawrence, and H. J. A. Röttgering (2011). A luminous quasar at a redshift of z = 7.085. *Nature 474*(7353), 616–9.

Ojala, T., M. Pietikäinen, and T. Mäenpää (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*, 971–987.

Perlmutter, S., G. Aldering, G. Goldhaber, R. A. Knop, P. Nugent, P. G. Castro, S. Deustua, S. Fabbro, A. Goobar, D. E. Groom, I. M. Hook, A. G. Kim, M. Y. Kim, J. C. Lee, N. J. Nunes, R. Pain, C. R. Pennypacker, R. Quimby, C. Lidman, R. S. Ellis, M. Irwin, R. G. McMahon, P. Ruiz-Lapuente, N. Walton, B. Schaefer, B. J. Boyle, A. V. Filippenko, T. Matheson, A. S. Fruchter, N. Panagia, H. J. M. Newberg, and W. J. Couch (1999). Measurements of Ω and Λ from 42 High-Redshift Supernovae. *The Astrophysical Journal 517*(2), 565–586.

Petrosian, V. (1976). Surface brightness and evolution of galaxies. *Astrophysical Journal Letters 209*, L1–L5.

Riess, A. G., A. V. Filippenko, P. Challis, A. Clocchiatti, A. Diercks, P. M. Garnavich, R. L. Gilliland, C. J. Hogan, S. Jha, R. P. Kirshner, B. Leibundgut, M. M. Phillips, D. Reiss, B. P. Schmidt, R. A. Schommer, R. C. Smith, J. Spyromilio, C. Stubbs, N. B. Suntzeff, and J. Tonry (1998). Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant. *The Astronomical Journal 116*(3), 1009–1038.

Schade, D., S. J. Lilly, D. Crampton, F. Hammer, O. Le Fevre, and L. Tresse (1995). Canada-France Redshift Survey: Hubble Space Telescope Imaging of High-Redshift Field Galaxies. *Astrophysical Journal Letters 451*, L1.

Sérsic, J. L. (1963). Influence of the atmospheric and instrumental dispersion on the brightness distribution in a galaxy. *Boletin de la Asociacion Argentina de Astronomia La Plata Argentina 6*, 41.

Skibba, R. A., S. P. Bamford, R. C. Nichol, C. J. Lintott, D. Andreescu, E. M. Edmondson, P. Murray, M. J. Raddick, K. Schawinski, A. Slosar, A. S. Szalay, D. Thomas, and J. Vandenberg (2009). Galaxy zoo: disentangling the environmental dependence of morphology and colour. *Monthly Notices of the Royal Astronomical Society 399*(2), 966–982.

Sporring, J. and J. Weickert (1999). Information measures in scale-spaces. *IEEE Transactions on Information Theory 45*(3), 1051–1058.

Steenstrup Pedersen, K., K. Stensbo-Smidt, A. Zirm, and C. Igel (2013). Shape index descriptors applied to texture-based galaxy analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2440–2447.

Stensbo-Smidt, K., F. Gieseke, C. Igel, A. Zirm, and K. Steenstrup Pedersen (2017). Sacrificing information for the greater good: how to select photometric bands for optimal accuracy. *Monthly Notices of the Royal Astronomical Society 464*(3), 2577–2596.

Stensbo-Smidt, K., C. Igel, A. Zirm, and K. S. Pedersen (2013). Nearest Neighbour Regression Outperforms Model-based Prediction of Specific Star Formation Rate. In *Proceedings of the IEEE International Conference on Big Data*, pp. 141–144. IEEE.

Tola, E., V. Lepetit, and P. Fua (2010). DAISY: An Efficient Dense Descriptor Applied to Wide-Baseline Stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*(5), 815–830.

Varma, M. and A. Zisserman (2005). A Statistical Approach to Texture Classification from Single Images. *International Journal of Computer Vision 62*(1/2), 61–81.

York, D. G., J. Adelman, J. E. Anderson, Jr., S. F. Anderson, J. Annis, N. A. Bahcall, J. A. Bakken, R. Barkhouser, S. Bastian, E. Berman, W. N. Boroski, S. Bracker, C. Briegel, J. W. Briggs, J. Brinkmann, R. Brunner, S. Burles, L. Carey, M. A. Carr, F. J. Castander, B. Chen, P. L. Colestock, A. J. Connolly, J. H. Crocker, I. Csabai, P. C. Czarapata, J. E. Davis, M. Doi, T. Dombeck, D. Eisenstein, N. Ellman, B. R. Elms, M. L. Evans, X. Fan, G. R. Federwitz,

L. Fiscelli, S. Friedman, J. A. Frieman, M. Fukugita, B. Gillespie, J. E. Gunn, V. K. Gurbani, E. de Haas, M. Haldeman, F. H. Harris, J. Hayes, T. M. Heckman, G. S. Hennessy, R. B. Hindsley, S. Holm, D. J. Holmgren, C.-h. Huang, C. Hull, D. Husby, S.-I. Ichikawa, T. Ichikawa, Ž. Ivezić, S. Kent, R. S. J. Kim, E. Kinney, M. Klaene, A. N. Kleinman, S. Kleinman, G. R. Knapp, J. Korienek, R. G. Kron, P. Z. Kunszt, D. Q. Lamb, B. Lee, R. F. Leger, S. Limmongkol, C. Lindenmeyer, D. C. Long, C. Loomis, J. Loveday, R. Lucinio, R. H. Lupton, B. MacKinnon, E. J. Mannery, P. M. Mantsch, B. Margon, P. McGehee, T. A. McKay, A. Meiksin, A. Merelli, D. G. Monet, J. A. Munn, V. K. Narayanan, T. Nash, E. Neilsen, R. Neswold, H. J. Newberg, R. C. Nichol, T. Nicinski, M. Nonino, N. Okada, S. Okamura, J. P. Ostriker, R. Owen, A. G. Pauls, J. Peoples, R. L. Peterson, D. Petravick, J. R. Pier, A. Pope, R. Pordes, A. Prosapio, R. Rechenmacher, T. R. Quinn, G. T. Richards, M. W. Richmond, C. H. Rivetta, C. M. Rockosi, K. Ruthmansdorfer, D. Sandford, D. J. Schlegel, D. P. Schneider, M. Sekiguchi, G. Sergey, K. Shimasaku, W. A. Siegmund, S. Smee, J. A. Smith, S. Snedden, R. Stone, C. Stoughton, M. A. Strauss, C. Stubbs, M. SubbaRao, A. S. Szalay, I. Szapudi, G. P. Szokoly, A. R. Thakar, C. Tremonti, D. L. Tucker, A. Uomoto, D. Vanden Berk, M. S. Vogeley, P. Waddell, S.-i. Wang, M. Watanabe, D. H. Weinberg, B. Yanny, N. Yasuda, and SDSS Collaboration (2000). The Sloan Digital Sky Survey: Technical Summary. *Astronomical Journal 120*, 1579–1587.