Mixture models for photometric redshifts

Zoe Ansari, Adriano Agnello, and Christa Gall

DARK, Niels Bohr Institute, University of Copenhagen, Jagtvej 128, 2200 Copenhagen, Denmark
e-mail: zakieh.ansari@nbi.ku.dk

Received 14 October 2020 / Accepted 2 March 2021

ABSTRACT

Context. Determining photometric redshifts (photo- z s) of extragalactic sources to a high accuracy is paramount to measure distances in wide-field cosmological experiments. With only photometric information at hand, photo- z s are prone to systematic uncertainties in the intervening extinction and the unknown underlying spectral-energy distribution of different astrophysical sources, leading to degeneracies in the modern machine learning algorithm that impacts the level of accuracy for photo- z estimates.

Aims. Here, we aim to resolve these model degeneracies and obtain a clear separation between intrinsic physical properties of astrophysical sources and extrinsic systematics. Furthermore, we aim to have meaningful estimates of the full photo- z probability distribution, and their uncertainties.

Methods. We performed a probabilistic photo- z determination using mixture density networks (MDN). The training data set is composed of optical (*griz* photometric bands) point-spread-function and model magnitudes and extinction measurements from the SDSS-DR15 and WISE mid-infrared ($3.4\,\mu\text{m}$ and $4.6\,\mu\text{m}$) model magnitudes. We used infinite Gaussian mixture models to classify the objects in our data set as stars, galaxies, or quasars, and to determine the number of MDN components to achieve optimal performance.

Results. The fraction of objects that are correctly split into the main classes of stars, galaxies, and quasars is 94%. Furthermore, our method improves the bias of photometric redshift estimation (i.e., the mean $\Delta z = (z_p - z_s)/(1 + z_s)$) by one order of magnitude compared to the SDSS photo- z , and it decreases the fraction of 3σ outliers (i.e., $3 \times \text{rms}(\Delta z) < \Delta z$). The relative, root-mean-square systematic uncertainty in our resulting photo- z s is down to 1.7% for benchmark samples of low-redshift galaxies ($z_s < 0.5$).

Conclusions. We have demonstrated the feasibility of machine-learning-based methods that produce full probability distributions for photo- z estimates with a performance that is competitive with state-of-the-art techniques. Our method can be applied to wide-field surveys where extinction can vary significantly across the sky and with sparse spectroscopic calibration samples. The code is publicly available.

Key words. methods: statistical – astronomical databases: miscellaneous – catalogs – surveys

1. Introduction

The redshift of an astrophysical object is routinely determined from absorption or emission lines in its spectrum. In the absence of spectroscopic information, its photometric redshift (hereafter photo- z) can be estimated from the apparent luminosity measured in different photometric bands (see e.g., Salvato et al. 2019, for a general review). Accurate photo- z s are needed by wide-field surveys that seek to probe cosmology through the spatial correlations of the matter density field, and are in fact a core limiting factor in the accuracy of these measurements (e.g., Knox et al. 2006).

While large areas of the sky are covered by optical and near-IR imaging surveys, only a minority of objects have observed spectra, and hence secure redshifts from emission or absorption lines. The major problem is the rather narrow wavelength range covered by most photometric bands that introduces uncertainties and degeneracies in estimating the redshift. The Kilo-Degree Survey (KiDS; de Jong et al. 2013) and Dark Energy Survey (DES; Abbott et al. 2018) collaborations have used a combination of spectroscopic surveys to calibrate photo- z s (e.g., Hoyle et al. 2018; Joudaki et al. 2018), with the ultimate aim of measuring the matter content (Ω_m) and present-day root-mean-square (rms) matter density fluctuations (σ_8). The most used spectroscopic datasets are zCOSMOS (Lilly & Zcosmos Team

2008), PRIMUS¹ (Coil et al. 2011), VVDS (Garilli et al. 2008), DEEP2 (Newman et al. 2013), VIPERS (Guzzo et al. 2014), GAMA (Driver et al. 2011), and SDSS (Ahn et al. 2012). Albeit to a different extent, their completeness is generally affected by a non-trivial pre-selection in colour and morphology, and in some cases a limited footprint and depth. Hildebrandt et al. (2020) have identified the different calibrations of photo- z s, across DES and KiDS, to explain the difference in inferred cosmological parameters, claiming that the uncertainties in photo- z s are one outstanding challenge towards percent-level cosmology from weak lensing.

When only photometric information is available, a three-fold degeneracy between an object type, its redshift, and foreground extinction hinders the unambiguous determination of the redshift. Galametz et al. (2017) have quantified this effect explicitly in view of a possible synergy between the ESA-*Euclid* mission (Amiaux et al. 2019) and *Rubin*-Legacy Survey of Space and Time (LSST; Amiaux et al. 2019), which should cover more than half of the extragalactic sky to $\gtrsim 24$ mag depth in *YJH*-bands and *ugriz*-bands, respectively. Here, we explore a probabilistic approach to compute photo- z s that account for

¹ While it is not all present in the literature, PRIMUS has been used for instance by Hoyle et al. (2018), and Behroozi et al. (2019) suggest that its redshift uncertainties may be smaller than previously thought.

the existence of an indefinite number of astrophysical object types and their cross-contamination due to broad-band imaging information. Specifically, we trained a suite of mixture density networks (MDNs, Bishop 1994) to predict the probability distribution of the photo- z of an object with measured magnitudes in multiple photometric bands as well as Galactic extinction. Following the standard nomenclature of machine-learning works, we alternatively refer to the photometric properties (magnitudes and extinction) as features in the rest of this paper. The MDN output is a sum of Gaussian functions in photo- z , whose parameters (i.e., the average, dispersion, amplitude) are non-linear combinations of the photometric inputs such as magnitude and extinction. Throughout the paper, we refer to these output Gaussians as ‘branches’. In order to determine the number of branches that are needed to optimally parameterise the photo- z probabilities, we must determine the range of MDN branches that most accurately describe the data set. Hence, we explore infinite Gaussian mixture models (IGMM) on a photometric sample of which about 2% of the sources have spectroscopic redshifts (see Sect. 2.1).

There are two main methods commonly used to estimate photometric redshifts: (i) template fitting and (ii) machine learning algorithms. Template fitting methods specify the relation between synthetic magnitudes and redshift with a suite of spectral templates across a range of redshifts and object classes, through maximum likelihood (e.g., Fernández-Soto et al. 1999) or Bayesian techniques (e.g., Benítez 2000; Brammer et al. 2008; Ilbert et al. 2006). Machine learning methods, using either images or a vector of magnitudes and colours, learn the relation between magnitude and redshift from a training data set of objects with known spectroscopic redshifts. In principle, template fitting techniques do not require a large sample of objects with spectroscopic redshifts for training, and can be applied to different surveys and redshift coverage. However, these methods are computationally intensive and require explicit assumptions on, for example, dust extinction, which can lead to a degeneracy in colour-redshift space. Moreover, template fitting techniques are only as predictive as the family of available templates. In the case of large samples of objects with spectroscopic redshifts, machine learning approaches such as artificial neural networks (ANNs; e.g., Amaro et al. 2019; Shuntov et al. 2020), k -nearest neighbours (kNN; e.g., Curran 2020; Graham et al. 2018; Nishizawa et al. 2020), tree-based algorithms (e.g., Carrasco Kind & Brunner 2013; Gerdes et al. 2010) or Gaussian processes (e.g., Almosallam et al. 2016) have shown similar or better performances than the template fitting methods. However, machine learning algorithms are only reliable in the range of input values of their training data set. Additionally, a lack of sufficient high redshift spectroscopic samples affects the performance of machine learning implementations on photo- z estimates. Another aspect is the production of photo- z probability distributions given the photometric measurements: While template-based methods can easily produce a probability distribution by combining likelihoods from different object templates, most of the machine-learning methods in the literature are only trained to produce point estimates, that is just one photo- z value for each object. For the sake of completeness, we summarise the state-of-the-art (and heterogeneous) efforts in the literature in Table A.1 and their performance metrics evaluation in Table A.2. We emphasise that most of the photo- z estimation methods above have been trained and tested purely on spectroscopic samples of different types of galaxies, often in a limited redshift range. Additionally, some of the spectroscopic galaxy samples were simulated entirely.

Here, we explore different kinds of mixture models to produce appropriate photo- z probability distributions that naturally account for the superposition of multiple, a priori unknown classes of astrophysical objects (e.g., stars, galaxies, quasars). There are multiple ways to describe a distribution of such objects in photometry space that consists of, for instance, magnitudes and extinction estimates (see Sect. 2.1) and that is also termed ‘feature space’ following the standard machine-learning terminology.

First, we use an IGMM (Teh 2010) to separate the astrophysical objects in feature space. This approach allows the algorithm to cluster the objects based on all the available photometric information without forcing the algorithm to classify the objects in a pre-determined way. Subsequently, the structure of the photometric (feature) space defines the number of Gaussian mixture components. Whenever a spectroscopic subsample of different types of astrophysical objects is available, IGMMs allow to separate this sample into classes, ideally representing each type of object. Secondly, we train MDNs to predict the photo- z probability distributions of objects in our data set. To find the optimal results, we explore different MDN implementations, which all include the IGMM components and membership probabilities obtained in the first step next to the entire photometric (feature) space (Sect. 2.1).

In Sect. 2, we describe our chosen training and test data sets as well as the IGMM and MDN implementations. The obtained accuracy of the classification along with the precision of the inferred photo- z s are provided in Sect. 3. In Sect. 5 we discuss our results, shortcomings and future improvements on our photo- z estimation alongside a comparison with other methods to estimate photo- z s from the literature.

2. Data and methods

To train our machine learning algorithms, we require a data set that contains: (i) morphological information from publicly available object catalogues (e.g., psf vs. model magnitudes, or stellerity index), to aid the separation of stars from galaxies and quasars; (ii) a wide footprint of the sky, to cover regions with sufficiently different extinction; (iii) multiband photometry from optical to mid-IR wavelengths, possibly including u -band; and (iv) a spectroscopic sub-sample of different types of objects (here: stars, galaxies and quasars).

2.1. Data

Our photometric data set is composed of optical PSF and model $griz$ -band magnitudes including i -band extinction measurements from the SDSS-DR15 (Aguado et al. 2019). We combine these SDSS magnitudes with $w1mpro$ and $w2mpro$ magnitudes (hereafter $W1$, $W2$) from WISE (Wright et al. 2010). We query the data in CasJobs² on the PhotoObjA11 table with a SDSS-WISE cross-match, requiring magnitude errors lower than 0.3 mag and $i - W1 < 8$ mag. Adding $g - r$, $r - i$, $i - z$, $z - W1$ and $W1 - W2$ colours leaves us with 22 dimensions to be used by our MDNs. However, the colours are strictly speaking redundant as they are obtained from the same, individual photometric bands. While this introduces many null value Eigenvectors in the IGMM, additional combinations of measurements are enabled, which speeds up the MDN computations by detrending the magnitude-magnitude distribution. Our spectroscopic data set (from SDSS-DR15) includes only objects with uncertainties on

² <https://skyserver.sdss.org/casjobs/>

Fig. 1. Spectroscopic data set in equatorial coordinates. Data are taken from SDSS-DR15 + WISE totalling about 245 000 objects of which there are 86 412 stars (yellow), 83 119 galaxies (purple) and 75 955 quasars (green). The entire photometric data set is a sample of about 1 023 000 objects, of which 98% lack spectroscopic redshifts and classification.

Fig. 2. Histogram showing i -band magnitudes of the objects from the photometric (blue) and spectroscopic data sets for stars (yellow), galaxies (purple) and quasars (green), in 0.1 magnitude bins.

Fig. 3. Redshift distribution of the spectroscopic data set. *Top panel:* galaxies (purple) and quasars (green) in 0.1 redshift bins. *Bottom panel:* stars (yellow) in 0.0001 redshift bins.

their spectroscopic redshift (from the SDSS pipelines) smaller than 1%. For only one MDN training, we added u -band PSF as well as model magnitudes. Our complete data sets are composed of a photometric and a spectroscopic data set. For about 2% of the photometric data set we have spectroscopic information. This data set is called the ‘spectroscopic data set’. For the IGMM, we have in total 1 022 731 unique sources from PhotoObjA11 and WISE, with additional 11 358 unique galaxies from WiggleZ (Drinkwater et al. 2010) cross-matched with PhotoObjA11 and WISE. For the test samples, the spectroscopic data set contains 86 412 unique stars, 83 119 unique galaxies and 75 955 quasars from SpecPhoto and WISE according to the classification of their spectra by the SDSS pipelines (see Fig. 1).

Figure 2 shows a general issue that is common to the literature on photo- z s, i.e., (by survey construction) the spectroscopic training sets do not reach the same depth as the photometric ones. This highlights the need for techniques that can extrapolate smoothly and with realistic uncertainties outside the ranges of a limited spectroscopic training set. Figure 3 shows the redshift distribution of different classes of objects from the spectroscopic data set. Evidently, galaxies are mainly placed at redshifts $\lesssim 1$, while quasars extend out to redshifts ~ 7 .

2.2. Infinite Gaussian mixture models

In a Gaussian mixture model (GMM), the density distribution of objects in ‘feature space’ (equivalent to photometric space, see Sect. 2.1) is described by a sum of Gaussian density components. The GMM is a probabilistic model which requires that a data set is drawn from a mixture of Gaussian density functions. Each Gaussian distribution is called a component. As the Gaussian distributions are defined in all the dimensions of the feature space, they are characterised by a mean vector and a covariance matrix. The feature vector contains the photometric information of each astronomical source. To describe the GMM, whenever needed, we use the notation $\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$, where $k \in \{1, \dots, K\}$ is the component index, μ_k , Σ_k and π_k are the mean vector and the covariance matrix in feature space, and the weight of component k , respectively.

Since the GMM is a Bayesian method, it requires multiple sets of model parameters and hyperparameters. The model parameters (means, covariances) change across the Gaussian components, while the hyperparameters are common to all of the Gaussian components, because they describe the priors from which all Gaussian components are drawn. For the GMM, the number of Gaussian components is a fixed hyperparameter.

Fig. 4. Maximum number of components vs. final number of components for different IGMM realisations, restricted to Gaussian components that contain at least 0.5% of the photometric data. Blue filled circles represent IGMM realisations that needed more than 2000 iterations to converge, while purple filled circles mark IGMM realisations that needed less than 2000 iterations. The size of the symbols scales with three different values of the prior of the Dirichlet concentration (γ). The light blue shaded region represents the confidence interval of 99% of regression estimation over the IGMM profiles by a multivariate smoothing procedure.

The IGMM is the GMM case with an undefined number of components, which is optimised by the model itself, depending on the photometric data set used. In particular, the IGMM describes a mixture of Gaussian distributions on the data population with an infinite (countable) number of components, using a Dirichlet process (Teh 2010) to define a distribution on the component weights.

However, setting an initial number of Gaussian density components is required by the IGMM. Based on the weights that are given to each such component at the end of the model training, it is common practice to exclude the least weighted components and define the data population only by the highest-weighted components. To pursue a fully Bayesian approach, it is advisable to explore a set of model hyperparameters with different initial guesses for the number of components. Like its finite GMM counterpart, each realisation of IGMM estimates the membership probability of each data point to each component. Appendix B provides a summary of the IGMM formalism.

For this work, we used the built-in variational IGMM package from the `scikit-learn` library for our implementations. In practice, the variational optimiser uses a truncated distribution over component weights with a fixed maximum number of components, known as stick-breaking representation (Ferguson 1973), with an expectation-maximisation algorithm (Dempster et al. 1977). To optimise the model and find the best representation of the data set, we explored a range of parameters. We increased the maximum number of allowed Gaussian components from 10 to 100 in increments of 2 and we set the maximum number of iterations for expectation maximisation performance to 2000. We used a Dirichlet concentration (γ) of each Gaussian component (k) on the weight distribution, of either 0.01, 0.05 or 0.0001 times the number of objects in the training data set. The covariance matrix for each Gaussian component was defined as type ‘full’; as per definition, each component has its own general covariance matrix. Furthermore, the prior on the mean distribution for each Gaussian component is defined as the median of the entries of the input vectors of the training data set (i.e., magnitudes, extinction).

Whenever needed, each object is assigned to the component to which its membership probability is maximal. In that case, we say that a component contains a data point.

The IGMM provides different possible representations of the same data set for each set of hyperparameters: here, we are interested in finding out the optimal number of components that can adequately describe the majority of the data. We then introduce a lower threshold on the number of sources that each component contains, and drop the components which contain less than the threshold. The threshold is defined by considering the size of the photometric sample and the highest value that we considered for the Dirichlet γ prior. The IGMM starts with components that contribute to 0.5% of the size of the photometric sample, since the highest γ value is 510 000 (see appendix for further details), due to our chosen ranges of hyperparameters. Therefore, we use 0.5% of the size of the photometric data set as the threshold. Figure 4 shows that the final number of components converges to 48 ± 4 . The convergence indicates that the models do not need more than 48 ± 4 components to describe the sample. Moreover, the initial 1:1 ramp-up in the figure shows that the final number of components is the same as the maximum tolerance, and so the model cannot adequately describe the data set; this trend breaks at about 44 components. To guide the eye, we determine a regression surface of all the IGMM profiles by a multivariate smoothing procedure³. In what follows, we choose 52 components.

The first IGMM implementation was fully unsupervised, as it was optimised to only describe the distribution of the objects in feature space. Subsequently, we trained different IGMMs considering additional spectroscopic information available for $\approx 2\%$ of the photometric sample. In particular, these partially supervised implementations are trained using the entire photometric feature space including either (I) spectroscopic classifications or (II) spectroscopic redshifts or (III) spectroscopic classifications and redshifts. Since the objects with additional spectroscopic information are a small part of the photometric training sample ($\approx 2\%$), the implementations ensure that the SDSS spectroscopic preselection does not bias the IGMM over the entire photometric sample. Finally, we calculate the membership probabilities to the 52 components for each object in the spectroscopic data set ($\approx 2.45 \times 10^5$ objects) from the optimised IGMM. This allows us to assign each object from the spectroscopic sample to one component. Thereafter, we label each of the IGMM components based on the percentage of spectroscopic classes that it contains.

Figure 5 shows the population of objects from the spectroscopic data set and their corresponding IGMM components in $g-r$ vs. $z-w1$ (upper panel) and $w2$ vs. $w1-w2$ (bottom panel) colour-colour and colour-magnitude diagrams. Each row from left to right shows the assigned components to stars, galaxies and quasars in the respective panels.

2.3. Mixture density networks

MDNs are a form of ANNs, which are capable of arbitrarily accurate approximation to a function and its derivatives based on the Universal Approximation Theorem (Hornik 1991). ANNs can be used for regression or classification purposes. ANNs are structured in layers of neurons, where each neuron receives an input vector from the previous layer, and outputs a nonlinear function of it that is passed on to the next layer. In MDNs, the aim is to approximate a distribution in the product space of input vectors of the individual sources (f_i) and target values (e.g.,

³ <https://has2k1.github.io/scikit-misc/stable/generated/skmisc.loess.loess.html>

Fig. 5. Colour-colour and colour-magnitude diagrams. Shown are $g - r$ vs. $z - W1$ colour-colour diagrams (*upper panel*) and $W2$ vs. $W1 - W2$ colour-magnitude diagrams (*bottom panel*) for a populations of objects from the spectroscopic data set such as stars (*left column*), galaxies (*middle column*) and quasars (*right column*). The purple contours correspond to the 68th percentile of each Gaussian IGMM component. The green filled circles correspond to the means μ_k of the Gaussian components. The grey scale indicates the abundance of the sources in each diagram.

$z_{s,i}$) as a superposition of different components. MDNs (Bishop 1994) are trained to optimise the log-likelihood

$$\log \mathcal{L} = \sum_{i=1}^N \log \left(\sum_{k=1}^{N_c} \hat{p}_k(f_i) \mathcal{N}(z_{s,i} | m_k(f_i), s_k(f_i)) \right) \quad (1)$$

by approximating the averages $m_k(f)$, amplitudes $\hat{p}_k(f)$ and widths $s_k(f)$. Here, N is the number of objects in the spectroscopic data set, while N_c denotes the number of output components (or branches) of the MDN.

Due to the limited information provided by the photometric space, a source of a specific spectroscopic class and low redshift can be confused with a different spectroscopic class and high redshift. Therefore, by providing distributions over a full range of redshifts, MDNs can cope with the fact that colours are not necessarily monotonic with redshift (as is the case e.g., in quasars). In order to avoid confusing MDN components with IGMM components, here we call MDN components branches.

For the sake of reproducibility, we use a publicly available MDN wrapper around the keras ANN module⁴ and a simple MDN architecture. The MDN input layer contains the same

photometric features (see Sect. 2.1) along with the membership probabilities of the IGMM, which carry additional information of the object classes (stars, galaxies and quasars). The dimension of the MDN input space is 74, of which 52 are the IGMM membership probabilities and 22 are the feature-space entries. The output layer of the MDN is defined by three neurons for each branch: the average redshift on the branch, the width of the branch and the membership probability of the source to one layer. The MDN is fully connected, that is the neurons in one layer are connected to all of the neurons in the next layer. Due to the fact that the MDN input contains the IGMM membership probabilities, after MDN hyperparameter optimisation, we train one MDN for each of the four IGMM implementations as described in previous sections.

We randomly split the entire spectroscopic data set Sect. 2.1 and use 80% for training and 20% for validation of the MDN. In order to optimise the MDN, we explored neural networks with 0 to 3 hidden layers. Each layer with a discrete number of neurons in the interval [3, 7, 10, 74, 100, 156, 222, 300, 400, 500, 528, 600, 740]. Furthermore, we used a discrete number of branches for the MDN in the interval [10, 52, 56, 100, 300]. The standard rectified linear unit (ReLU, Nair & Hinton 2010) and parametric rectified linear unit (PReLU, He et al. 2015) are used as

⁴ <https://github.com/cmpmpercussion/keras-mdn-layer>

Fig. 6. MDN Loss ($-\log(\mathcal{L})/N$) as a function of epoch. The loss obtained during the MDN training and validation are shown by blue and orange lines, respectively.

Fig. 7. IGMM confusion matrix. The spectroscopic classifications are shown against the IGMM classes of the spectroscopic data set.

activation functions. We used ADAM as optimiser and batch learning with 64 objects per epoch with learning rates of 10^{-6} , 10^{-5} , 10^{-4} and 10^{-3} to mitigate local minima of the loss function.

By comparing the training and validation loss of MDNs with the previously defined set of hyperparameters, the resulting optimal set of hyperparameters contains a hidden layer with 528 neurons and 10 MDN branches. The activation function is PReLU and the learning rate for the ADAM optimiser is 10^{-4} . Figure 6 shows the loss function, $-\log(\mathcal{L})/N$, for the training and validation data set, for the MDN optimisation for which membership probabilities are obtained from the partially supervised IGMM realisation that also considers the spectroscopic classes. As Fig. 6 shows, the learning curve flattens roughly around 300 epochs. To mitigate overfitting, we concluded that 300 epochs are sufficient to train the model. Additionally to training MDNs with the redshifts as targets, we tested $\log(z_s)$ as a target and it led to an improvement in the z_p estimation.

3. Results

We trained an IGMM on the photometric data set (see Sect. 2.1), using the optimal hyperparameters (Sect. 2.2). Thereafter, we linked IGMM components to the three spectroscopic classes using a spectroscopic data set (Sect. 2.1). Finally, we implemented MDNs on the spectroscopic data set using photometric features and membership probabilities from the IGMM to estimate the conditional probability distribution $p(z_p|f)$ of photo- z values from the photometric inputs. In this section, we describe the evaluation methods and the resulting classification and photo- z estimations.

3.1. Classification

With our mixture models we address the common problem of cross contamination among different classes of objects due to the a priori unknown underlying spectral energy distribution. In the IGMM realisations, each object can belong to each of the components with a probability $p_{i,k} = w_k \mathcal{N}(f_i|\mu_k, \Sigma_k) / \sum_l (w_l \mathcal{N}(f_i|\mu_l, \Sigma_l))$, which we denote by membership probabilities in the following. As we introduced above (end of Sect. 2.2), the simplest way to assign an object (with feature vector f_i) to a component is to consider the component index $\hat{k}$ for which $p_{i,\hat{k}}$ is maximised.

To parameterise the accuracy of the classification, we consider the usual quantification of true and false positives and true and false negatives (e.g., Fawcett 2006), and build a confusion matrix to quantify the rate of correct classifications. Figure 7 shows the confusion matrix of the GMM-based classification for the spectroscopic data set. The true positive rates⁵ for stars, galaxies and quasars are 0.97, 0.91 and 0.95, respectively. False positive rates for stars that are true galaxies and quasars are 0.0029 and 0.029. False negative rates for stars that are assigned to galaxies and quasars are 0.031 and 0.019 of all stars, respectively. The accuracy⁶ is $\approx 94\%$. This means that the IGMM part of our mixture models can clean an extragalactic sample from most of stellar contaminants, and broadly separate galaxies from AGN-dominated objects. We compare the performance of our classification to that of a HDBSCAN classification (Campello et al. 2013) of a subset of 39 447 objects from the SDSS-DR14 that have an accuracy of $\approx 95\%$ using *ugrizYHJK*_{w1w2}-band magnitudes (Logan & Fotopoulou 2020). We find that our method achieves a comparable accuracy even without *u*-band and *YJHK*-band magnitudes.

Figure 5 demonstrates that the IGMM recognises the main behaviours of stars, galaxies and quasars in colour space and also identifies subclasses that are not highly represented in the spectroscopic sample, such as white dwarfs and brown dwarfs. On the other hand, some components happen to lie in regions of the colour-magnitude-extinction space that are not dominated by only one subclass. The overlap between different object classes in photometry can affect the classification performance and the output of the classification that is then used by the MDN regression. The components corresponding to regions of overlap between different classes are discussed below.

Approximately 30% of IGMM components that cover $\approx 15\%$ of the spectroscopic data set, marked in red in Table 1, contain a non-negligible fraction of objects from more than one of the three main classes. Figure 8 shows their position in the

⁵ Defined as: $TP/(TP+FN)$.

⁶ Defined as: $(TP+TN)/(TP+TN+FN+FP)$.

Table 1. Percentage of objects from each spectroscopic class (stars, galaxies, quasars) within each IGMM component.

IGMM components	Stars	Galaxies	Quasars
1	85.42	0.27	14.31
2	99.98	0	0.02
3	97.46	0.06	2.48
4	100	0	0
5	1.57	0.54	97.88
6	99.86	0.05	0.1
7	3.7	86.06	10.24
8	100	0	0
9	97.45	0.05	2.5
10	8.94	71.67	19.39
11	1.97	90.16	7.87
12	6.95	52.25	40.80
13	99.6	0	0.4
14	100	0	0
15	42.38	43.22	14.41
16	55.39	0.43	44.18
17	99.93	0.01	0.06
18	96.75	2.48	0.77
19	6.58	36.44	56.98
20	99.89	0	0.11
21	1.14	94.51	4.35
22	98.02	0.07	1.90
23	99.94	0	0.06
24	3.69	89.54	6.77
25	100	0	0
26	99.94	0.01	0.05
27	97.48	0.47	2.05
28	100	0	0
29	12.31	20.04	67.65
30	100	0	0
31	1.02	96.60	2.38
32	11.13	35.58	53.28
33	99.96	0.02	0.02
34	99.71	0	0.29
35	100	0	0
36	99.8	0.1	0.1
37	34.23	42.05	23.72
38	100	0	0
39	8.43	51.74	39.83
40	99.91	0	0.09
41	99.51	0.04	0.45
42	100	0	0
43	4.43	88.61	6.97
44	0.56	98.18	1.25
45	90.3	0.83	8.87
46	79.57	1.22	19.21
47	2.87	65.41	31.72
48	0.73	0.05	99.21
49	100	0	0
50	60.24	0.74	39.02
51	44.52	26.33	29.15
52	95.64	0.04	4.32

Notes. The components highlighted in red lie between different spectroscopic class regions in photometric feature space, and can reduce the classification accuracy.

same colour-colour and colour-magnitude diagrams as Fig. 5. We address these components as ‘problematic components’.

As expected, the problematic components lie at the faint end (with higher magnitude uncertainties in WISE), or in intermediate regions of the colour space between AGN-dominated and galaxy-dominated systems. Additionally, the SDSS spectroscopic classification of some objects is ambiguous and for some cases the automatic classification (by the SDSS spectral pipelines) is either erroneous or has multiple incompatible entries⁷. These issues occur more frequent for fainter objects which have spectra with low signal-to-noise ratio⁸. However, since most of the objects are clustered in three main classes which are correctly identified by the IGMM components, uncertain spectroscopic labels are not a significant problem for our calculations.

3.2. Photometric redshifts

Here we discuss different metrics employed to evaluate the performance of our methods used to determine photometric redshifts. Most metrics are based on commonly used statistical methods as outlined. The prediction bias is defined as the mean of weighted residuals, $\Delta z = (z_p - z_s)/(1 + z_s)$, which is the same as the definition in Cohen et al. (2000). The root-mean-square of the weighted residuals is defined as $\text{rms}(\Delta z)$. The fraction of outliers is defined as the number of objects with $3 \times \text{rms}(\Delta z) < \Delta z$.

For all methods, we excluded objects with spectroscopic redshift errors $\delta z_s > 0.01 \times (1 + z_s)$. For each source, the MDN determines a full photo- z distribution, which is a superposition of all branches, each with a membership probability, average, and dispersion. If one so-called point estimate is needed, there are at least two options to compute it. One option is the expectation value

$$\mathbb{E}(z_{p,i} | f_i) = \frac{\sum_k \mu_k(f_i) \hat{p}_k}{\sum_k \hat{p}_k}, \quad (2)$$

weighted across all branches that an object can belong to, according to its branch membership probabilities. Another choice, which we follow here, is the peak $\mu_r(f_i)$ of the branch that gives the maximum membership probability of a given object. In what follows, we refer to this redshift value as the ‘peak photo- z ’.

In a fully Bayesian framework, one would also consider the maximum-a-posteriori, which in our approach would correspond with the photo- z that maximises the MDN output sum of Gaussians because all our priors are uniform. Strictly speaking, in principle the maximum-a-posteriori and the peak redshift are not the same, if the MDN Gaussian with the highest membership probability lies close to other MDN Gaussians. Here, for the sake of computational convenience we choose the peak redshift, leaving the comparison with other possible choices (including the maximum-a-posteriori, or non-uniform priors) to future investigation. We also note that this distinction becomes important for objects whose membership probabilities are not clearly dominated by one of the MDN output Gaussians. In the following, we examine the photo- z estimation performance on objects whose membership probabilities to one of the MDN Gaussians is $> 80\%$, in which case there is almost no difference between peak redshift and maximum-a-posteriori redshift.

⁷ For example for OBJID=1691188859137714176 from SDSS-DR15.

⁸ For example for OBJID=743142903307593728 from SDSS-DR15.

Fig. 8. Colour-colour and colour-magnitude diagrams. Shown are $g-r$ vs. $z-W1$ colour-colour diagrams (*upper panel*) and $W2$ vs. $W1-W2$ colour-magnitude diagrams (*bottom panel*) for objects from the spectroscopic data set of the three spectroscopic classes such as stars (*left column*), galaxies (*middle column*) and quasars (*right column*). The purple contours correspond to the 68-th percentile of the problematic Gaussian components of the IGMM that are not dominated by objects of just one spectroscopic class. The green filled circles correspond to the means μ_k of these components. The grey scale indicates the number of sources in each diagram.

Figure 9 shows the distribution of peak photo- z s (top) and expectation photo- z s (bottom) versus spectroscopic redshifts, z_s , for the MDN run with ten branches. One aspect to consider when determining photo- z in cosmological wide-field imaging surveys, is the availability of u -band magnitudes, which is currently available for KiDS but not for DES. The *Rubin* LSST is expected to deliver u -band photometry at the same depth of KiDS over $\approx 12\,300\text{deg}^2$ (Y1) and $\approx 14\,300\text{deg}^2$ (Y2) (The LSST Dark Energy Science Collaboration 2018). To test the effect, we re-trained one of our mixture models (IGMM spec. class) for a data set that includes u -band PSF and model magnitudes as additional input features (Fig. 10). The bias and root-mean-square residuals are provided in Table 2 for all objects and for galaxies with spectroscopic redshifts $z_s < 0.3$, $z_s < 0.4$, and $z_s < 0.5$. This test leads to a lower rms Δz and smaller fraction of 3σ outliers than for the same model without u -band magnitudes and can be considered an improvement in accuracy. One reason is that for low redshift objects, u -band contains information on the position and strength of the $4000\text{\AA}$ Balmer-break, which is not covered by other, longer wavelength bands. Furthermore, with respect to the cross-contamination problem, this model also improves the overall confidence level with which an

object belongs to a branch. As demonstrated in Fig. 10, bottom panel, the MDN performs indeed better for objects with increased confidence level.

4. Conclusion

Performance evaluations of MDN photo- z s from different IGMM realisations are summarised in Table 2. The minimum MDN photo- z bias and rms of different IGMM realisations without using u -band magnitudes is reached by the IGMM realisation (Spec. class, (z_s)) that uses the spectroscopic classifications and spectroscopic redshifts of $\approx 2\%$ of the dataset. This implies that the IGMM can provide a reasonable description of the objects in their feature space even with very limited spectroscopic information. On the other hand, for this realisation the fraction of 3σ outliers is not as low as for other IGMM realisations (Tables 2 and 3) which have a slightly higher bias and rms. For the validation samples that are restricted to galaxies with a spectroscopic redshift < 0.3 , the minimum MDN photo- z bias, rms and fraction of 3σ outliers are achieved for the realisation that includes u -band magnitudes. The MDN performance of different IGMM realisations for objects that belong to one MDN branch with high

Fig. 9. Comparison of spectroscopic vs. MDN photometric redshifts. The photometric redshifts are taken from the partially supervised ‘spec. class’ IGMM implementation (as described in Sect. 2.2). The colour-scales indicate the number of objects. *Top panels:* predicted photometric redshifts that correspond to the branches with the highest weights. The single panels show the weights, dispersions (denoted by ‘width’) and residuals *from top to bottom*. *Bottom panels:* mean photometric redshifts of the predicted redshifts over all branches with respect to their weights. *Lower panel:* residuals. *Left panels:* include all classes with $z_{\text{spec}} < 7$. *Right panels:* include all galaxies with $z_{\text{spec}} < 0.3$.

confidence (i.e., $\text{weight}_{\text{max}} > 0.8$) is summarised in Table 3. In order to compare our photo- z estimates to those from the SDSS, we select objects from the SDSS SpecObjAll table that have photo- z s (obtained with a kNN interpolation) from the SDSS PhotoObjAll (38 487 objects). We remark that our MDN yields the full photo- z PDF, so a choice must be made when comparing its results with point estimates from other methods in the literature. For this reason, we follow our previous choice and consider the peak photo- z ’s. We compare our photo- z ’s to the SDSS ones for the full sample (38 487 objects) and also for the subset of objects that belong to a MDN branch with high confidence ($\text{weight}_{\text{max}} > 0.8$, 18 355 objects). To account for uncertainties in bias and rms due to the finite sample size, we split the whole sample into five sub-sets, compute the bias and rms on each, and then report their average and standard deviation. As an alternative, we also evaluate the bias, rms and fraction of outliers when the MDN is trained with a k -fold cross-validation method (see e.g. Altman & Bland 2005) on the full SDSS training set ($\approx 245\,000$ objects), where each of the $k = 5$ folds exclude

a fifth of the SDSS photo- z sample. As summarised in Table 4 and shown in Fig. 11, our method improves the bias of photo- z estimates by about one order of magnitude compared to SDSS photo- z s for objects for which our model estimated the photo- z s at high confidence (i.e., $\text{weight}_{\text{max}} > 0.8$), and our method also decreases the rms and the fraction of outliers. All metrics are improved, with the added advantage that the MDN computes photo- z s for all objects (instead of just those with low stellarity) and can also cover the $z_s > 1$ range more accurately than the SDSS kNN. As a matter of fact, the SDSS photo- z s hardly exceed $z_p \approx 1$, while our machinery is trained over a much wider redshift range.

As a general benchmark, the LSST system science requirements document⁹ defines three photometric redshift requirements for a sample of four billion galaxies with $i < 25$ mag within $z_s < 0.3$ as follows: First, for the error in $(1+z_s)$ the rms of

⁹ <https://docushare.lsstcorp.org/docushare/dsweb/Get/LPM-17>

Fig. 10. Photo- z performance of different MDN implementations. *Top panel:* retaining only objects with $\text{weight}_{\text{max}} > 0.8$ membership probability to a MDN branch. *Middle panel:* including u -band PSF and model magnitudes. *Bottom panel:* u -band magnitudes and MDN branch $\text{weight}_{\text{max}} > 0.8$. *Left column:* all types of objects in the spectroscopic data set. *Right column:* only galaxies in $z_s < 0.3$. from the spectroscopic data set.

residuals is lower than 0.02. Secondly, the fraction of 3σ ('catastrophic') outliers is lower than 10% of the sample and thirdly, the bias is lower than 0.003.

In our approach, these requirements are met if the MDN peak z_p is adopted. The rms Δz can be brought to 0.02 over $0 < z_s < 0.5$ both for all objects (Table 5) and for 'high-confidence' objects with >0.8 membership probability to a branch (Table 6; called weight in Sect. 2.3). Although our training and evaluation

samples from the SDSS do not reach as deep as the *Rubin* LSST is expected to, our method shows promising results as a starting point. A general remark, which holds for all photo- z estimation methods is that a re-training for the LSST regime may also need deeper photometric and more robust spectroscopic samples. Recently, Beck et al. (2021) used neural networks to classify objects in the Pan-STARRS1 footprint, which is known to have a more accurate photometry than the SDSS

Table 2. MDN performance evaluation, without any clipping for the average and rms, without any threshold on branch membership probabilities.

IGMM implementation	Photometry	$\langle\Delta z\rangle$ (all)	rms(Δz) (all)	3σ outliers (all)	$\langle\Delta z\rangle$, range1 ^(a)	rms(Δz), range1 ^(a)	3σ outliers, range1 ^(a)	rms(Δz), range2 ^(b)	rms(Δz), range3 ^(c)
Fully unsup.	<i>griz</i> , W1, W2	0.0152	0.2174	3.08%	0.0007	0.0177	0.28%	0.0988	0.0945
Spec. class	<i>griz</i> , W1, W2	0.0111	0.2069	1.31%	0.0006	0.0167	0.41%	0.0822	0.0783
Spec. class ^(d)	<i>griz</i> , W1, W2	0.0356	0.2300	1.35%	0.0110	0.0260	0.71%	0.0953	0.0903
Redshift (z_s)	<i>griz</i> , W1, W2	0.0176	0.2131	3.21%	-0.0009	0.0174	0.38%	0.0896	0.0873
Spec. class, z_s	<i>griz</i> , W1, W2	0.0047	0.1990	2.66%	0.0036	0.0181	0.57%	0.0675	0.0664
Spec. class	<i>ugriz</i> , W1, W2	0.0135	0.1592	1.62%	0.0007	0.0160	0.23%	0.0601	0.0611

Notes. Spectroscopic sample for all IGMM implementations containing stars, galaxies and quasars. ^(a)Restricted to galaxies with $z_s < 0.3$; ^(b)restricted to galaxies with $z_s < 0.4$; ^(c)restricted to galaxies with $z_s < 0.5$; ^(d)expectation value. Fully unsup.: Fully unsupervised IGMM implementation. Spec. class: uses spectroscopically classified objects (e.g., stars, galaxies, and quasars) for those which SDSS provides spectroscopic information (i.e., $\approx 2\%$ of the photometric data set) to classify the data. Redshift (z_s): uses the spectroscopic redshift of objects from $\approx 2\%$ of the photometric data set to classify the data. Spec. class, (z_s): uses both, spectroscopically classified objects and the spectroscopic redshift of objects from $\approx 2\%$ of the photometric data set to classify the data.

Table 3. MDN performance evaluation exclusively for sources with MDN branch weight_{max} > 0.8, without any clipping for the average and rms.

IGMM implementation	Photometry	$\langle\Delta z\rangle$ (all)	rms(Δz) (all)	3σ outliers (all)	$\langle\Delta z\rangle$, range1 ^(a)	rms(Δz), range1 ^(a)	3σ outliers, range1 ^(a)	rms(Δz), range2 ^(b)	rms(Δz), range3 ^(c)
Fully unsup.	<i>griz</i> , W1, W2	0.0032	0.1165	1.00%	0.0007	0.0177	0.60%	0.0350	0.0360
Spec. class	<i>griz</i> , W1, W2	0.0031	0.1244	0.93%	0.0006	0.0167	0.83%	0.0405	0.0391
Redshift (z_s)	<i>griz</i> , W1, W2	0.0035	0.1076	0.79%	-0.0009	0.0174	0.52%	0.0299	0.0331
Spec. class, z_s	<i>griz</i> , W1, W2	-0.0048	0.1170	1.02%	0.0036	0.0036	1.02%	0.0337	0.0314
Spec. class	<i>ugriz</i> , W1, W2	0.0043	0.0934	0.66%	0.0007	0.0160	0.92%	0.0334	0.0341

Notes. ^(a)Restricted to galaxies with $z_s < 0.3$; ^(b)restricted to galaxies with $z_s < 0.4$; ^(c)restricted to galaxies with $z_s < 0.5$. The IGMM implementation abbreviations are the same as for Table 2.

Table 4. Comparison between the photo- z evaluation on all objects from the spectroscopic sample with available SDSS photo- z s.

No k-fold in MDN; IGMM implementation	Restriction on objects	Number of objects	Bias	rms	3σ outliers
SDSS <i>griz</i> photo- z	None	38 632	0.0001 ± 0.0008	0.0458 ± 0.0483	$(0.73 \pm 0.22)\%$
MDN spec. class + <i>griz</i> , W1, W2	None	38 632	0.0018 ± 0.0014	0.0478 ± 0.0384	$(0.49 \pm 0.04)\%$
SDSS <i>griz</i> photo- z	weight _{max} > 0.8	18 355	-0.0038 ± 0.0040	0.0394 ± 0.0413	$(0.46 \pm 0.19)\%$
MDN spec. class + <i>griz</i> , W1, W2	weight _{max} > 0.8	18 355	-0.0002 ± 0.0009	0.0389 ± 0.0318	$(0.55 \pm 0.26)\%$
k-fold in MDN; IGMM implementation	Restriction on objects	Number of objects	Bias	rms	3σ outliers
SDSS <i>griz</i> photo- z	None	38 632	0.0001 ± 0.0020	0.0662 ± 0.0072	$(0.79 \pm 0.22)\%$
MDN spec. class + <i>griz</i> , W1, W2	None	38 632	0.0027 ± 0.0031	0.0724 ± 0.0139	$(0.45 \pm 0.10)\%$
SDSS <i>griz</i> photo- z	weight _{max} > 0.8	23 908	-0.0028 ± 0.0029	0.0506 ± 0.0096	$(0.33 \pm 0.14)\%$
MDN spec. class + <i>griz</i> , W1, W2	weight _{max} > 0.8	23 908	0.00011 ± 0.0042	0.0458 ± 0.0069	$(0.28 \pm 0.07)\%$

Table 5. MDN performance evaluation.

IGMM implementation	Photometry	$\langle\Delta z\rangle$ ($z_s < 1$)	rms(Δz) ($z_s < 1$)	$\langle\Delta z\rangle$, range1 ^(a)
Fully unsup.	<i>griz</i> , W1, W2	0.0007 ± 0.0006	0.0204 ± 0.0106	0.0025 ± 0.0005
Spec. class	<i>griz</i> , W1, W2	0.0030 ± 0.0028	0.0216 ± 0.0104	0.0042 ± 0.0039
Redshift (z_s)	<i>griz</i> , W1, W2	0.0022 ± 0.0024	0.0207 ± 0.0113	0.0008 ± 0.0003
Spec. class, z_s	<i>griz</i> , W1, W2	0.0001 ± 0.0005	0.0208 ± 0.0112	0.0058 ± 0.0006
Spec. class	<i>ugriz</i> , W1, W2	0.0023 ± 0.0026	0.0181 ± 0.0091	0.0004 ± 0.0025
IGMM implementation	photometry	rms(Δz), range1 ^(a)	rms(Δz), range2 ^(b)	rms(Δz), range3 ^(c)
Fully unsup.	<i>griz</i> , W1, W2	0.0215 ± 0.0013	0.0233 ± 0.0013	0.0238 ± 0.0015
Spec. class	<i>griz</i> , W1, W2	0.0230 ± 0.0015	0.0245 ± 0.0019	0.0247 ± 0.0014
Redshift (z_s)	<i>griz</i> , W1, W2	0.0219 ± 0.0014	0.0238 ± 0.0012	0.0242 ± 0.0013
Spec. class, z_s	<i>griz</i> , W1, W2	0.0218 ± 0.0019	0.0235 ± 0.0019	0.0235 ± 0.0024
Spec. class	<i>ugriz</i> , W1, W2	0.0179 ± 0.0015	0.0182 ± 0.0016	0.0187 ± 0.0019

Notes. The bias and rms are computed using the definition of clipped bias and rms in PS1-STR (Beck et al. 2021). ^(a)Restricted to galaxies with $z_s < 0.3$; ^(b)restricted to galaxies with $z_s < 0.4$; ^(c)restricted to galaxies with $z_s < 0.5$.

Table 6. MDN performance evaluation exclusively for sources with MDN branch $\text{weight}_{\text{max}} > 0.8$.

IGMM implementation	Photometry	$\langle \Delta z \rangle (z_s < 1)$	$\text{rms}(\Delta z) (z_s < 1)$	$\langle \Delta z \rangle, \text{range1}^{(a)}$
Fully unsup.	<i>griz</i> , W1, W2	0.0004 ± 0.0010	0.0249 ± 0.0038	0.0014 ± 0.0005
Spec. class	<i>griz</i> , W1, W2	0.0035 ± 0.0030	0.0273 ± 0.0045	0.0015 ± 0.0009
Redshift (z_s)	<i>griz</i> , W1, W2	0.0018 ± 0.0013	0.0236 ± 0.0045	-0.0004 ± 0.0004
Spec. class, z_s	<i>griz</i> , W1, W2	-0.0014 ± 0.0012	0.0264 ± 0.0054	0.0042 ± 0.0005
Spec. class	<i>ugriz</i> , W1, W2	0.0036 ± 0.0026	0.0227 ± 0.0058	0.0009 ± 0.0005
IGMM implementation	photometry	$\text{rms}(\Delta z), \text{range1}^{(a)}$	$\text{rms}(\Delta z), \text{range2}^{(b)}$	$\text{rms}(\Delta z), \text{range3}^{(c)}$
Fully unsup.	<i>griz</i> , W1, W2	0.0186 ± 0.0002	0.0201 ± 0.0003	0.0207 ± 0.0004
Spec. class	<i>griz</i> , W1, W2	0.0181 ± 0.0002	0.0207 ± 0.0009	0.0214 ± 0.0007
Redshift (z_s)	<i>griz</i> , W1, W2	0.0182 ± 0.0004	0.0195 ± 0.0005	0.0203 ± 0.0004
Spec. class, z_s	<i>griz</i> , W1, W2	0.0189 ± 0.0013	0.0205 ± 0.0014	0.0207 ± 0.0014
Spec. class	<i>ugriz</i> , W1, W2	0.0163 ± 0.0007	0.0167 ± 0.0008	0.0170 ± 0.0010

Notes. The bias and rms are computed using the definition of clipped bias and rms in PS1-STR (Beck et al. 2021). ^(a)Restricted to galaxies with $z_s < 0.3$; ^(b)restricted to galaxies with $z_s < 0.4$; ^(c)restricted to galaxies with $z_s < 0.5$.

(Magnier et al. 2013), and evaluated photo- z s on objects with a probability $p > 0.8$ of being galaxies, obtaining $\text{rms}(\Delta z) = 0.03$ over $0 < z_s < 1$. If we follow the same definitions and clipping¹⁰ as by Beck et al. (2021), then we obtain 1.7–2% relative rms over the $0 < z_s < 0.5$ redshift range. Our code for re-training and any further evaluation is publicly accessible¹¹.

5. Discussion

Adding u -band information, as is the case with the SDSS and will be the case with the LSST, reduces the bias and fraction of outliers in all the redshift ranges considered. This is also because adding u -band magnitudes sharpens the MDN separation into branches and increases the fraction of objects with the highest weighted branch > 0.8 , as can be seen in the bottom panels of Fig. 9.

We remark that throughout this work, we are simply adopting reddening values in the i -band (A_i), which the SDSS provides via a simple conversion of measured $E(B - V)$ values with a Milky-Way extinction law and $R_V = 3.1$. Our approach accounts for the systematic uncertainties due to the unknown extinction law by producing probability distributions and associate uncertainties for each photo- z value.

The combined information across the optical and infrared, through the SDSS and WISE magnitudes, helps reducing the overlap between different classes in colour-magnitude space. The WISE depth is not a major limiting factor in the sample completeness as long as samples from the SDSS are considered, but it can affect the completeness significantly for deeper surveys (Spiniello & Agnello 2019). In view of performing the classification and photo- z estimation on the DES, and on the Rubin LSST later on, deeper mid-IR data are needed. The unWISE reprocessing of the WISE cutouts improved upon the original WISE depth (Lang 2014). Further in the future, forced photometry of the unWISE cutouts from wide-field optical and NIR surveys may further increase the mid-IR survey depth (e.g., Lang & Hogg 2014).

In general, separating objects into many sub-classes aids the photo- z regression, as each MDN branch only needs to consider

Fig. 11. Top panel: SDSS spectroscopic redshift vs. SDSS photometric redshift. Bottom panel: spectroscopic redshift vs. photometric redshift (this work). Colour bars indicate the number of sources in the diagrams. The selection of sources is made by retaining objects with $\text{weight}_{\text{max}} > 0.8$ membership probability to a MDN branch.

a subset of objects with more homogeneous properties than the whole photometric sample. Furthermore, the approach that we used in this work is both in the realm of machine learning (hence less constrained by choices of templates) while it can also produce a full output distribution for the photo- z given the available photometric information. Beyond their first implementation in

¹⁰ Their clipping procedure removes objects with $|\Delta z| > 0.15$.

¹¹ github.com/ZoeAnsari/MixtureModelsForPhotometricRedshifts

this work, mixture models can be easily adapted so that they can account for missing entries and limited depth, as in the GMM implementation by [Melchior & Goulding \(2018\)](#).

Acknowledgements. This work is supported by a VILLUM FONDEN Investigator grant (project number 16599), a VILLUM FONDEN Young Investor Grant (project number 25501), and a Villum Experiment Grant (project number 36225). This project is partially funded by the Danish council for independent research under the project “Fundamentals of Dark Matter Structures”, DFF–6108-00570.

References

- Abbott, T. M. C., Abdalla, F. B., Allam, S., et al. 2018, *ApJS*, **239**, 18
- Aguado, D. S., Ahumada, R., Almeida, A., et al. 2019, *ApJS*, **240**, 23
- Ahn, C. P., Alexandroff, R., Allende Prieto, C., et al. 2012, *ApJS*, **203**, 21
- Almosallam, I. A., Lindsay, S. N., Jarvis, M. J., et al. 2016, *MNRAS*, **455**, 2387
- Altman, D. G., & Bland, J. M. 2005, *BMJ*, **331**, 903
- Amaro, V., Cavuoti, S., Brescia, M., et al. 2019, *MNRAS*, **482**, 3116
- Amiaux, J., Scaramella, R., Mellier, Y., et al. in *Space Telescopes and Instrumentation 2012: Optical, Infrared, and Millimeter Wave*, SPIE Conf. Ser., 8442, 84420Z
- Beck, R., Szapudi, I., Flewelling, H., et al. 2021, *MNRAS*, **500**, 1633
- Behroozi, P., Wechsler, R. H., Hearin, A. P., & Conroy, C. 2019, *MNRAS*, **488**, 3143
- Benítez, N. 2000, *ApJ*, **536**, 571
- Bishop, C. M. 1994, unpublished
- Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, *ApJ*, **686**, 1503
- Campello, R. J. G. B., Moulavi, D., & Sander, J. 2013, in *Advances in Knowledge Discovery and Data Mining*, eds. J. Pei, V. L. Cao, H. Motoda, & G. Xu (Berlin, Heidelberg: Springer), 160
- Carrasco Kind, M., & Brunner, R. J. 2013, *MNRAS*, **432**, 1483
- Cohen, J. G., Hogg, D. W., Blandford, R., et al. 2000, *ApJ*, **538**, 29
- Coil, A. L., Blanton, M. R., Burles, S. M., et al. 2011, *ApJ*, **741**, 8
- Curran, S. J. 2020, *MNRAS*, **493**, L70
- de Jong, J. T. A., Verdoes Kleijn, G. A., Kuijken, K. H., et al. 2013, *Exp. Astron.*, **35**, 25
- Dempster, A. P., Laird, N. M., & Rubin, D. B. 1977, *J. R. Stat. Soc.: Ser. B (Methodological)*, **39**, 1
- Drinkwater, M. J., Jurek, R. J., Blake, C., et al. 2010, *MNRAS*, **401**, 1429
- Driver, S. P., Hill, D. T., Kelvin, L. S., et al. 2011, *MNRAS*, **413**, 971
- Fawcett, T. 2006, *Pattern Recognit. Lett.*, **27**, 861
- Ferguson, T. S. 1973, *Ann. Statist.*, **1**, 209
- Fernández-Soto, A., Lanzetta, K. M., & Yahil, A. 1999, *ApJ*, **513**, 34
- Galamez, A., Saglia, R., Paltani, S., et al. 2017, *A&A*, **598**, A20
- Garilli, B., Le Fèvre, O., Guzzo, L., et al. 2008, *A&A*, **486**, 683
- Gerdes, D. W., Sypniewski, A. J., McKay, T. A., et al. 2010, *ApJ*, **715**, 823
- Görür, D., & Edward Rasmussen, C. 2010, *J. Comp. Sci. Technol.*, **25**, 653
- Graham, M. L., Connolly, A. J., Ivezić, Ž., et al. 2018, *AJ*, **155**, 1
- Guzzo, L., Scodreggio, M., Garilli, B., et al. 2014, *A&A*, **566**, A108
- He, K., Zhang, X., Ren, S., & Sun, J. 2015, ArXiv eprints [arXiv:1502.01852]
- Hildebrandt, H., Köhlinger, F., van den Busch, J. L., et al. 2020, *A&A*, **633**, A69
- Hornik, K. 1991, *Neural Networks*, **4**, 251
- Hoyle, B., Gruen, D., Bernstein, G. M., et al. 2018, *MNRAS*, **478**, 592
- Ilbert, O., Arnouts, S., McCracken, H. J., et al. 2006, *A&A*, **457**, 841
- Joudaki, S., Blake, C., Johnson, A., et al. 2018, *MNRAS*, **474**, 4894
- Knox, L., Song, Y.-S., & Zhan, H. 2006, *ApJ*, **652**, 857
- Lang, D. 2014, *AJ*, **147**, 108
- Lang, D., & Hogg, D. W. 2014, *ApJ*, **151**, 36
- Lilly, S., & Zcosmos Team, 2008, *The Messenger*, **134**, 35
- Logan, C. H. A., & Fotopoulou, S. 2020, *A&A*, **633**, A154
- Magnier, E. A., Schlafly, E., Finkbeiner, D., et al. 2013, *ApJS*, **205**, 20
- Melchior, P., & Goulding, A. D. 2018, *Astron. Comput.*, **25**, 183
- Nair, V., & Hinton, G. E. 2010, in *Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10* (Madison, WI, USA: Omnipress), 807
- Newman, J. A., Cooper, M. C., Davis, M., et al. 2013, *ApJS*, **208**, 5
- Nishizawa, A. J., Hsieh, B. C., Tanaka, M., & Takata, T. 2020, ArXiv eprints [arxiv:2003.01511]
- Pasquet, J., Bertin, E., Treyer, M., et al. 2019, *A&A*, **621**, A26
- Sadeh, I., Abdalla, F. B., & Lahav, O. 2019, *ANNz2: Estimating Photometric Redshift and Probability Density Functions Using Machine Learning Methods*
- Salvato, M., Ilbert, O., & Hoyle, B. 2019, *Nat. Astron.*, **3**, 212
- Schmidt, S. J., Malz, A. I., Soo, J. Y. H., et al. 2020, *MNRAS*, **499**, 1587
- Shuntov, M., Pasquet, J., Arnouts, S., et al. 2020, *A&A*, **636**, A90
- Spiniello, C., & Agnello, A. 2019, *A&A*, **630**, A146
- Teh, Y. W. 2010, in *Dirichlet Process*, eds. C. Sammut, & G. I. Webb (Boston, MA: Springer, US), 280
- The LSST Dark Energy Science Collaboration (Mandelbaum, R., et al.) 2018, ArXiv eprints [arXiv:1809.01669]
- Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, *AJ*, **140**, 1868

Appendix A: Summary of photometric redshifts in the literature

Table A.1. Recent automated approaches to estimate photo-zs.

Reference	Method ^(a)	Photometric information	Objects	z_s range ^(b)	Depth [mag] ^(c)	Survey
1	kNN	<i>ugrizy</i> ^(d)	Galaxies	$0 < z \leq 2$	$i < 25.3$	mock galaxy for LSST from DESC
2	ANN	<i>ugriz</i> ^(e) , $E(B - V)$	Galaxies	$z < 0.4$	$r_{\text{Petro}} \leq 17.8$	SDSS-DR12
3	METAPHOR, ANN, template fitting	<i>ugri</i> GAaP	Galaxies	$z_s \leq 1$	$r \leq 21$	SDSS-DR9, KiDS ESO-DR3, GAMA-DR2, 2dFGRS
4	ANN	<i>ugriz</i> ^(e)	Galaxies	$z_s \leq 0.4$	$r_{\text{Petro}} \leq 17.8$	SDSS/BOSS-DR12, GAMA-DR3
5	kNN	<i>UV, ugrizy, YJHK</i>	Galaxies	$0.3 < z_s < 3.0$	$i < 25$	Mock galaxy catalogues for <i>Euclid</i> , RST, and/or CASTOR
6	kNN	<i>UV, ugriz, w1w2w3w4</i>	QSOs	–	$14.7 < r < 22.6$	SDSS-DR12, 2MASS, WISE
7	kNN	<i>grizy</i>	Galaxies	$z_s > 0.01$	$18.5 < i < 25$	SDSS/BOSS-DR14, DEEP2/3DR4, VANDELs-DR2, COSMOS, C3R2, COSMOS2015
8	Tree based	<i>ugriz, BRI</i>	Galaxies	$0.02 \leq z_s \leq 0.3$	$B_{AB} < 24.1$	SDSS/MGS-DR7, DEEP2-DR4
9	Tree based	<i>ugriz</i>	Galaxies	$z_s \leq 0.55$	$r_{\text{Petro}} < 17.77$	SDSS-DR6, 2dF-SDSS LRG, 2SLAQ, DEEP2
10	Gaussian process	<i>griz, RIZ, YJH</i>	Galaxies	$0 \leq z_s \leq 2$	$RIZ < 25$	SDSS/BOSS
11	ensemble of ANNs, trees and kNN	<i>ugriz</i>	Galaxies	$z_s < 0.8$	$i_{AB} \lesssim 22.5$	SDSS/BOSS-DR10
12	ANN, Monte-Carlo, extrapolation	<i>grizy</i> ^(f) , $E(B - V)$ ^(g)	Galaxies, QSOs, Stars	$z_s < 1.5$	$i \lesssim 23.1$	PS1 3 π DR1, SDSS-DR14, DEEP2-DR4, VIPERS PDR-2, WiggleZ, zCOSMOS-DR3, VVDS

Notes. ^(a) Acronyms are defined in the respective literature; ^(b) spectroscopic redshift range. ^(c) Petrosian r -band magnitude, r_{Petro} ; ^(d) grey scale 48×48 pixel images; ^(e) images in *ugriz*, 64×64 pixels in each band; ^(f) magnitudes for PSF, Kron and seeing-matched apertures (FPSFMag, FKronMag and FApMag, respectively), as well as $3.00''$, $4.63''$ and $7.43''$ fixed-radius apertures (FmeanMagR5, FmeanMagR6 and FmeanMagR7); ^(g) PS1 and *Planck* extinction maps.

References. (1) Schmidt et al. (2020); (2) Pasquet et al. (2019); (3) Amaro et al. (2019); (4) Shuntov et al. (2020); (5) Graham et al. (2018); (6) Curran (2020); (7) Nishizawa et al. (2020); (8) Carrasco Kind & Brunner (2013); (9) Gerdes et al. (2010); (10) Almosallam et al. (2016); (11) Sadeh et al. (2019); (12) Beck et al. (2021).

Table A.2. Comparison of photo-z estimates.

Reference	Method ^(a)	Bias ^(b)	rms ^(c)	Fraction of outlier in %
1	(trainZ)	−0.2086	0.1808	0
	ANNz2	0.00063	0.0270	4.4
	BPZ	−0.00175	0.0215	3.5
	Delight	−0.00185	0.0212	3.8
	EAZY	−0.00218	0.0225	3.4
	FlexZBoost	−0.00027	0.0154	2.0
	GPz	0.00000	0.0197	5.2
	Lephare	−0.00161	0.0236	5.8
	METAPhoR	0.00000	0.0264	3.7
	CMNN	−0.00132	0.0184	3.5
	SkyNet	−0.00167	0.0219	3.6
	TPZ	0.00309	0.0161	3.3
2	Convolutional neural network(CNN)	0.0001	0.0456 ^(d)	0.31
3	METAPHOR	−0.004	0.065	0.98
	ANNz2	−0.008	0.078	1.60
	BPZ	−0.020	0.048	1.13
4	CNN + density field (mode)	0.0038 ^(d)		0.83
	CNN + density field (median)	0.0045 ^(d)	–	0.44
	CNN + density field (mean)	0.0066 ^(d)		0.31
5	kNN	−0.0001 ± 0.0	0.0165 ± 0.0001	4.0
6	kNN	0.001 ^(e)	0.36	10.7 ^(f)
7	DEmP ^(g)	−0.0291	0.1018	0.16
	DEmP ^(h)	−0.0175	0.07	0.17
8	Trees and random forest(Regression mode)	−0.00008	0.0225	0
	Trees and random forest (Classification mode)	0.00218	0.0246	0
9	ArborZ	−0.006 ^(e)	0.985	1.9
10	GP-GL	0.0946	0.1420	5.3
	GP-VL	0.828	0.1251	5.5
	GP-VC	0.0294	0.0435	4.7
11	Ensemble of ANNs, trees and KNN (nominal solution)	0.0002	0.034	0.105
	Ensemble of ANNs, trees and KNN(<i>(PDF)</i>)	0.00035	0.034	0.105
	Ensemble of ANNs, trees and KNN(PDF)	0.00035	0.052	0.1
12	PS1-STRM (All validation) base estimate	0.0003	0.0342	2.88 ⁽ⁱ⁾
	PS1-STRM (All validation) Monte-Carlo sampled	0.0010	0.0344	2.99
	PS1-STRM (Non-extrapolated) base estimate	0.0005	0.0322	1.89
	PS1-STRM (Non-extrapolated) Monte-Carlo sampled	0.0013	0.0323	2.00

Notes. Values are provided where information was available. ^(a)Acronyms are defined in the respective literature; ^(b)Bias: defined as mean of $\Delta z = (z_p - z_s)/(1 + z_s)$; ^(c)rms($(z_p - z_s)/(1 + z_s)$); ^(d) $\sigma_{\text{MAD}} = 1.4826 \times \text{MAD}$, where MAD (Median Absolute Deviation) is the median of $|\Delta z - \text{Median}(\Delta z)|$; ^(e)Average of $\delta z = z_p - z_s$; ^(f)fraction of outliers defined as number of objects with $|\Delta z| > \text{rms}(\Delta z) \pm 0.5$; ^(g)exclusively using wide-band photometry from Wide fields of HSC (<https://hsc.mtk.nao.ac.jp/ssp/>) as additional photometric input; ^(h)exclusively using deep photometry from Deep and UltraDeep fields of HSC as additional photometric input; ⁽ⁱ⁾fraction of outliers defined as number of objects with $|\Delta z| > 0.15$.

References. (1) Schmidt et al. (2020); (2) Pasquet et al. (2019); (3) Amaro et al. (2019); (4) Shuntov et al. (2020); (5) Graham et al. (2018); (6) Curran (2020); (7) Nishizawa et al. (2020); (8) Carrasco Kind & Brunner (2013); (9) Gerdes et al. (2010); (10) Almosallam et al. (2016); (11) Sadeh et al. (2019); (12) Beck et al. (2021).

Appendix B: IGMM

Probability density distribution (PDF) formalisation by Gaussian mixture modelling for K components is defined as follows:

$$P(x|\mu_1, \dots, \mu_K, \Sigma_1, \dots, \Sigma_K) = \sum_{k=1}^K \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \quad (\text{B.1})$$

where x is the data, π_k is the weight distribution of mixtures that is defined by a Dirichlet distribution and $\sum_{k=1}^K \pi_k = 1$.

IGMM is the GMM case with infinite number of components using Dirichlet process instead of Dirichlet distribution to define the prior over the mixture distribution. Dirichlet process is a distribution over distributions, parameterising by concentration parameter γ and a base distribution G_0 . The base distribution is the Dirichlet distribution which is a prior over the locations of components in the parameter space (i.e., $\Theta = (\mu, \Sigma)$). The concentration parameter γ expresses the strength of belief in G_0 and affects the components weight (Görür & Edward Rasmussen 2010).

Based on Bayes rule:

$$\gamma_{Z_i}(k) = P(Z_i = k|x) = \frac{P(k)P(x|Z_i = k)}{P(x)} = \frac{\pi_k \mathcal{N}(x|\Theta_k)}{\sum_{k=1}^K \pi_k \mathcal{N}(x|\Theta_k)} \quad (\text{B.2})$$

where π is considered as the Dirichlet process and Z_i is the latent variable. $\pi_k = N_k/N$ represents the effective number of data points assigned to the k -th mixture component. Despite the

fact that we do not know the latent variable, there is information about it in the posterior.

Using an expectation-maximisation (EM) algorithm to find the maximum likelihood with respect to the model parameters includes two steps, estimation step (e-step) and maximisation step (m-step). After initialising the model parameters and evaluating the log-likelihood, the e-step evaluates the posterior distribution of Z_i using the current model parameter values by Eq. (B.2). Then, the m-step updates the model parameters based on the calculated latent variable as follows:

$$\mu_k = \frac{\sum_{i=1}^N \gamma_{Z_i}(k) x_i}{\sum_{i=1}^N \gamma_{Z_i}(k)} = \frac{1}{N_k} \sum_{i=1}^N \gamma_{Z_i}(k) x_i \quad (\text{B.3})$$

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^N \gamma_{Z_i}(k) (x_i - \mu_k)(x_i - \mu_k) \quad (\text{B.4})$$

$$\pi_k = \frac{N_k}{N} \quad (\text{B.5})$$

Where $N_k = \sum_{i=1}^N \gamma_{Z_i}(k)$. Eventually, the algorithm detects the convergence by the lack of significant change in the log-likelihood value from one iteration to the next, using:

$$\log P(x|\mu, \Sigma, \pi) = \sum_{i=1}^N \log \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k) \right\} \quad (\text{B.6})$$

where π_k , the mixture proportion, represents the probability of x_i belonging to the k -th mixture component.