Astroinformatics
Proceedings IAU Symposium No. 325, 2016
M. Brescia, S.G. Djorgovski, E. Feigelson,
G. Longo & S. Cavuoti, eds.

doi:10.1017/S1743921316013089

Dealing with Uncertain Multimodal Photometric Redshift Estimations

Kai L. Polsterer[1] †

[1] Astroinformatics, Heidelberg Institute for Theoretical Studies,
Schloss-Wolfsbrunnenweg 35, 69118, Heidelberg, Germany
email: kai.polsterer@h-its.org

Abstract. Due to limitations in available instrumentation and observation time, a spectroscopic determination of distance is not feasible for all objects in the sky. Therefore statistical methods that estimate redshifts, based on photometric measurement are of tremendous importance to many astrophysical questions. Determining cosmological parameters and understanding evolutionary processes in the universe are just two examples. When perform astrophysical analyses, it is necessary to treat the uncertainties of the estimates correctly. Over-simplification of results and the usage of wrong tools to evaluate the performance of probabilistic redshift estimates were commonly found in the literature. We present proper tools for evaluating uncertain redshift estimates and discuss the necessity of multimodal redshift distributions.

Keywords. instrumentation: miscellaneous, methods: data analysis, methods: miscellaneous, methods: statistical, techniques: photometric, galaxies: distances and redshifts

1. Introduction

Photometric redshift estimation is a regression problem where the redshift of an object should be derived from a set of low resolution photometric measurements. It is more efficient to derive photometric measurements instead of performing a full spectroscopic analysis. Multiple objects can be observed simultaneously with a much better signal-to-noise-ratio instead of being individually dispersed. Thereby bandpass filters of different widths can be combined to roughly reconstruct the spectral energy distribution of the objects and hence provide a basis for estimating the redshift of the object. The estimation of photometric redshift is done with different approaches. Some approaches are based on fitting a set of template spectra via minimizing the residuals between the observed photometric measurements and the integrated template fluxes. Other approaches make use of a very large set of photometric references and estimate the redshift with methods from the field of statistical learning. A different class of models tries to generalize the regression problem and uses machine learning techniques to learn a relation between the photometric input values and the redshift values.

In order to describe our continuous way of improving photometric redshift estimation techniques, the following subsections present our individual steps in a historically ordered listing.

1.1. *Improving Methods*

In the past, we started working on optimizing the photometric redshift estimation models. To improve the quality of photometric redshift estimates, we tried different approaches. We compared the performance of generalizing models and models that are local in the

† KLP would like to thank the Klaus Tschira Foundation for supporting this research.

space of reference objects (Gieseke *et al.* 2012). In this publication we learned how important the ability to generalize is, in case you have a severe shift between your reference data and the data, you want to apply your model to. Domain shift is a serious problem in the field of photometric redshift estimation, especially when dealing with high-redshift quasars. The limited number of good references is a problem for methods that are not generalizing. Therefore in Polsterer *et al.*(2013) we constrained the scientific question on detecting quasar that exceed a certain redshift limit. By using a limiting redshift of $z > 4.8$ instead of demanding accurate photometric redshift estimates, we could make use of a simple nearest neighbour model. This kind of model is limited to the reference values in the vicinity with respect to the high-dimensional feature space and therefore does not generalize. Due to the extreme size of the available reference data of spectroscopically confirmed quasars, this is an efficient and computationally not too demanding model to filter a very large data-set.

To further improve the performance of our photometric redshift estimation, we applied different machine learning techniques and dealt with different reference data-sets. Besides using random forest based approaches, we employed artificial neural networks and support vector regression models. We observed severe differences in the results obtained, when comparing them with other approaches in the literature. Even when using the same reference data-sets, differences could be observed. E.g. when estimating the photometric redshifts based on the quasars of Schneider *et al.*(2010), differences are introduced by interpreting the photometric flags individually, rejecting quasars and therefore changing the quality of the reference sample. The community should start working on generating a reference data-set that can be used for evaluating the performance of estimation approaches. Besides a non homogeneous set of quality measures, another source of differences in the results are unused cross-validations techniques and missing uncertainty quantification for the quality measurements. In Section 2 we introduce a set of proper tools and measures that should become standard when dealing with uncertain photometric redshift estimations.

1.2. *Performing Feature Selection*

A detailed analysis of the performances of the different approaches lead us to the optimization of the selection of the input features. We started with standard colors from neighboring filter bands. Back then, it could be observed in the literature that some other approaches started to incorporate the errors of the feature extraction mechanisms. Therefore we started adding more and more features, too. At a certain point, we could not understand how further improvements were achieved by adding more and more features. Especially, as some of the features like model- or petrosian-magnitudes had no physical relation to the redshift estimation task of quasars. Besides the used algorithms and techniques, the used features and feature-combinations are considerably contributing to the performance of the methods, too.

To further investigate the impact of feature combinations on the prediction performance, we started working on a systematic testing of feature combinations. In Heinermann *et al.*(2013), Polsterer *et al.*(2014) and Gieseke *et al.*(2014) we made used of fast graphical processing units to explore which kind of feature combinations help to improve the performance of the predictions. By employing features that are typically not used for photometric redshift estimation and creating all possible colors between these raw features, we could further improve our prediction performance by more than 25 percent.

https://doi.org/10.1017/S1743921316013089 Published online by Cambridge University Press

1.3. *Adding Uncertainty*

When improving the selection of input features for the regression models, we ended up using the measurement errors of the feature extraction pipeline. Those errors are reflecting the uncertainty of the measured magnitudes in the different filter bands. A first approach by us to generate output uncertainties was to apply sampling techniques. By sampling over the given variances of the input features and statistically combining the individually retrieved point estimates, very simple probability density functions (*PDF*s) were generated. In order to evaluate the performance of the photometric redshift probability distributions, proper tools have been analyzed and studied. The tools we finally ended up using are described in Section 2.

We found that sampling over the input uncertainties had just minor effect on the distribution of the output uncertainties. Due to the fact that a simple model was used that is trained to generate point estimates, the main source of model uncertainty was ignored. The uncertainty and the degeneracies that are introduced when recovering a complex spectral energy distribution via a few broad-band filters, are the dominating source of prediction uncertainty. By applying a proper model that has the flexibility to produce more complex predictive distributions, a significant improvement of the quality of the estimates could be achieved. In the next sections we discuss the necessity of complex *PDF*s as uncertain redshift estimates in more details.

Outline: This work is structured as follows: In Section 2 we present a set of proper tools to evaluate and inspect probabilistic redshift estimates. Next, we motivate the necessity of multimodal redshift distributions in Section 3. After presenting some experiments and results in Section 4 we conclude this work in Section 5.

2. Proper Tools

We provided a detailed description of proper measures and tools to evaluate the performance of uncertain redshift estimates in Polsterer *et al.*(2016). Therefore we put a copy of the important parts from this publication in this section.

As stated by Gneiting *et al.*(2007), when comparing forecasting distributions and observations, the goal is to maximize the sharpness of the predictive distributions subject to calibration. In the context of photometric redshift estimation this refers to comparing *PDF*s with spectroscopic redshifts. The term calibration describes the consistency between the predictive distribution and the true redshift. Sharpness is used to express the concentration of the *PDF*.

2.1. *Probability Integral Transform*

In Dawid (1984) the probability integral transform (*PIT*) was proposed to be used as a diagnostic tool to check the calibration and the sharpness of the generated predictive distributions. The *PIT* is a visual tool which is based on the histogram of the values of the cumulative probability at the true value. Therefore the *PDF*s have to be transferred into cumulative distribution functions (*CDF*s) (see Equation 2.1).

$$CDF_t(z_t) = \int_{-\infty}^{z_t} PDF_t(z)dz \tag{2.1}$$

With respect to photometric redshift estimations, the *PIT* is calculated with the *CDF* of the estimated redshift CDF_t at the true redshift z_t (see Equation 2.2). Hereby $t \in \{1, 2, \dots N\}$ indexes the corresponding tuple of a predicted redshift distribution and the matching true redshift for N data items.

https://doi.org/10.1017/S1743921316013089 Published online by Cambridge University Press

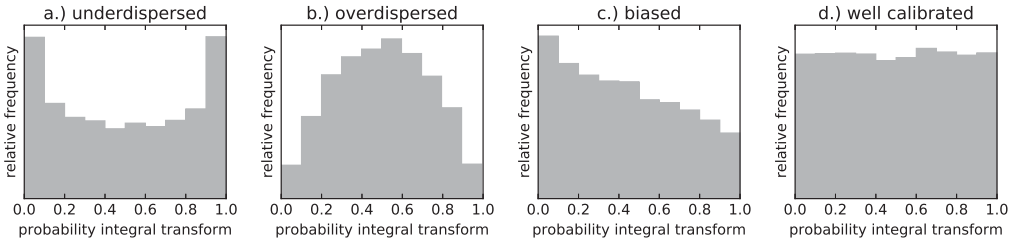

Figure 1. Four different probability integral transforms (*PIT*s). In the case of underdispersed *PDF*s an u-shaped, concave distribution is observed (a). Overdispersed *PDF*s result in a peaked, convex distribution (b). When a slope in the *PIT* is observed, the analysed *PDF*s are biased (c). Only when the *PIT* exhibits a flat distribution, the *PDF*s are well calibrated (d).

$$p_t = CDF_t(z_t) \tag{2.2}$$

When the anaylsed *PDF*s are of Gaussian nature with μ and σ^2 as mean and variance, the *CDF*s can be evaluated by using Equation 2.3. For a Gaussian mixture model, the corresponding *CDF* is a additive mixture of single Gaussian *CDF*s multiplied with their weights, respectively.

$$CDF_t(z_t) = \frac{1}{2}\left[1 + erf\left(\frac{z_t - \mu}{\sqrt{2\sigma^2}}\right)\right] \tag{2.3}$$

In case the predictions are optimal, the distribution of $p_t, t \in \{1, 2, \ldots N\}$ has to be uniform. As shown in Figure 1, multiple aspects can be verified by plotting the histogram of this distribution. Only if the distribution of p_t exhibits a uniform shape, the *PDF*s are well calibrated. When the dispersion of the estimates is to small in relation to the distribution of the true redshifts, an underdispersed distribution of p_t can be observed. This will be reflected by a u-shaped, concave histogram. The opposite case is observed with overdispersed *PDF*s that generate a peaked, concave histogram. As soon as a bias is present in the *PDF*s, a slope is added to the distribution of p_t.

The histogram of the *PIT* values allows to visually check the calibration and sharpness, i.e. testing how well the distribution of the p_t values is of uniform nature. It provides intuitive access to multiple aspects of the *PDF*s with respect to the corresponding true redshifts. Therefore we recommend this tool to be always used when evaluating *PDF*s.

2.2. *Continuous Ranked Probability Score*

When comparing different approaches that generated *PDF*s based on photometric features, a proper score should be used to measure the individual prediction performances. In this context, prediction performance refers to how well the generated *PDF*s represents the true redshifts of the objects in a test data-set. Please see Gneiting *et al.*(2007) for a detailed introduction to the topic of proper scoring rules. In this work we make use of the continuous ranked probability score (*CRPS*) as a performance measure. The *CRPS* (Hersbach 2000) is widely used in the field of weather forecasting for expressing a distance between a *PDF* and a true value. It compares a full distribution with an observation as defined in:

https://doi.org/10.1017/S1743921316013089 Published online by Cambridge University Press

$$CRPS = \frac{1}{N}\sum_{t=1}^{N} crps(CDF_t, z_t),$$
$$\text{with } crps(CDF_t, z_t) = \int_{-\infty}^{+\infty} \left[CDF_t(z) - CDF_{z_t}(z)\right]^2 dz \tag{2.4}$$

CDF_t is the cumulative distribution of the *PDF*, as defined in Equation 2.1. In Equation 2.5 the cumulative distribution of the true redshift CDF_{z_t} is defined based on $H(z) = \mathcal{H}$, the Heaviside step-function.

$$CDF_{z_t}(z) = H(z - z_t), \text{ with } H(z) = \begin{cases} 0 & \text{for } z < 0 \\ 1 & \text{for } z \geqslant 0 \end{cases} \tag{2.5}$$

In case the *PDF*s are given as normal distributions, we are able to write it in the subsequent form (Gneiting *et al.* 2005).

$$crps[\mathcal{N}(\mu_t, \sigma_t^2), z_t] = \sigma_t \left\{ \frac{z_t - \mu_t}{\sigma_t} \left[2\Phi\left(\frac{z_t - \mu_t}{\sigma_t}\right) - 1 \right] + 2\phi\left(\frac{z_t - \mu_t}{\sigma_t}\right) - \frac{1}{\sqrt{\pi}} \right\}, \tag{2.6}$$

where ϕ and Φ represent the *PDF* and the *CDF* of a normal distribution with mean 0 and variance 1, respectively. In Equation 2.6 the $\frac{z_t - \mu_t}{\sigma_t}$ term represents the normalized prediction error. Representing a *PDF* as a Gaussian mixture model (*GMM*) (Bishop 2007) provides some advantages in calculating the *CRPS* for even more complicated distribution. A Gaussian mixture model (see Equation 2.7) defines a distribution as a combination of M number of Gaussians with independent means μ and variances σ^2.

$$GMM(\mu, \sigma^2, \omega) = \sum_{i=1}^{M} \omega_i \mathcal{N}(\mu_i, \sigma_i^2),$$
$$\text{with } \sum_{i=1}^{M} \omega_i = 1 \text{ and } \omega_i \geqslant 0, \forall i \in \{1, 2, \ldots M\} \tag{2.7}$$

Hereby the weights ω control the contributions of the individual Gaussians to the final distribution. With

$$A(\mu, \sigma^2) = 2\sigma\phi\left(\frac{\mu}{\sigma}\right) + \mu\left[2\Phi\left(\frac{\mu}{\sigma}\right) - 1\right] \text{ and} \tag{2.8}$$

$$crps(GMM_t(\mu, \sigma^2, \omega), z_t) = \sum_{i=1}^{M} \omega_i * A(z_t - \mu_i, \sigma_i^2) - \sum_{i=1}^{M}\sum_{j=1}^{M} \frac{1}{2}\omega_i\omega_j * A(\mu_i - \mu_j, \sigma_i^2 - \sigma_j^2) \tag{2.9}$$

we can calculate the *CRPS* of a *GMM* (Grimit *et al.* 2006). Besides *GMM*s there are many other mixture models that could be used to represent the *PDF*s. For a lot of scores the `properscoring` package provides a *Python* implementation that can be easily used for calculating the *CRPS*. This package includes the calculation of the *CRPS* for an ensemble of predictions, too.

https://doi.org/10.1017/S1743921316013089 Published online by Cambridge University Press

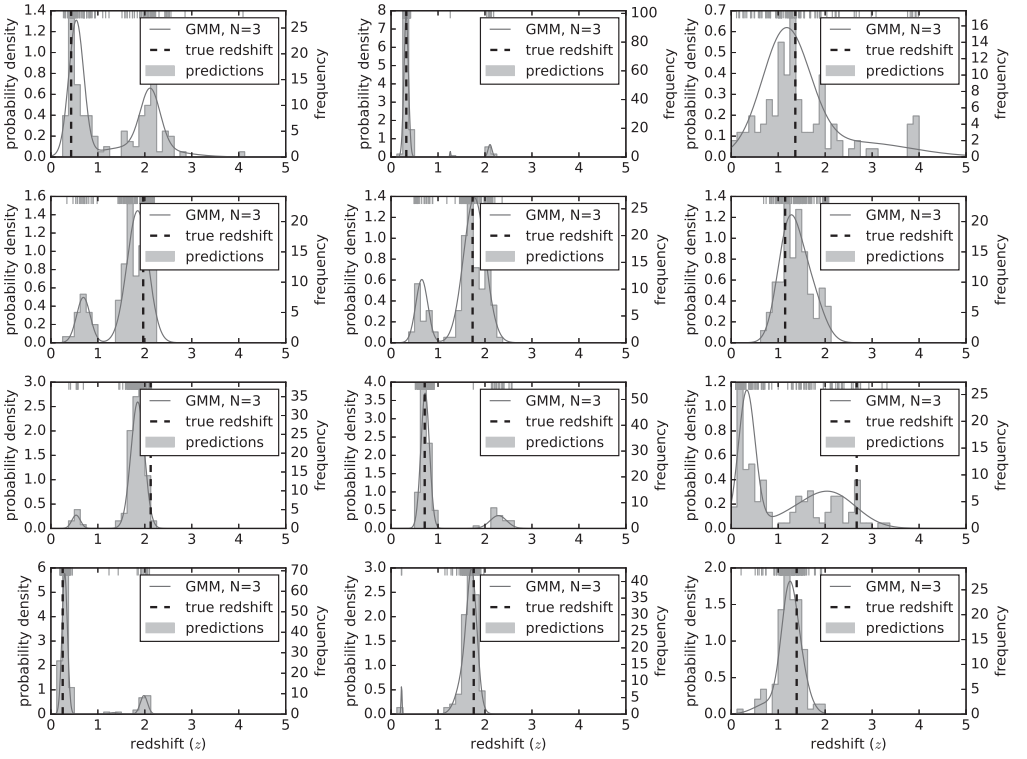

Figure 2. Example of 12 randomly chosen quasars from *SDSS* (*DR7*). The individual nearest neighbors are plotted at the top together with the histogram in the background. In addition, a Gaussian mixture model with three components and the true redshift are marked. A large fraction of the randomly selected objects show a complex or clear multimodal nature.

3. Multimodal Redshift Distributions

In order to understand the uncertainties that are introduced through the degeneracies of observing the complex spectral energy distribution of an object through a very broad and limited set of photometric filters, we did a very simple analysis. To be able to perfectly predict the true redshifts based on photometric measurements, there needs to be a bidirectional unique relation between the photometric input features and the spectroscopic redshifts. Under the assumption, that object that appear to be similar in the photometric filters should have similar spectroscopic redshift, as set of objects was randomly selected. The distribution of the true redshifts of the neighbors with respect to their distribution in the feature space was analyzed. As shown in Figure 2 our assumption is wrong. For each object we used the 128 nearest neighbors with respect to the distance in the feature space. Most of the objects show a clear multimodal distribution of the redshift values of the neighbouring objects. This distribution has no correlation to the distance in the feature space and therefore clearly indicates that our assumption is wrong.

In the case of photometric redshift estimation of quasars based on the broadband filters of *SDSS* there exists no unique relation between input and output values. This explains, why objects can be, e.q. equally likely at two different redshifts and a simple point estimate is not able to capture the whole problem. The fitted Gaussian mixture models gives a much smoother and a continuous representation of the probability densities. In addition, this representation makes it easier to compute and process the estimated *PDF*s.

https://doi.org/10.1017/S1743921316013089 Published online by Cambridge University Press

Producing single outputs with an uncertainty and dealing with unimodal distributions is not a correct approach to add uncertainties to photometric redshift estimations.

4. Experiments

As motivated above, photometric redshift estimation requires a more complex description of the probability distribution, than single point estimated with symmetric uncertainties could deliver. In this section we perform experiments that visualize the good prediction performances that can be achieved when using multimodal descriptions for the estimates.

4.1. *Reference Data and Data Preprocessing*

The data for our experiments is taken from *SDSS* (*DR7*). Based on the quasar catalog by Schneider *et al.*(2010), we extracted 80,000 objects together with their *ugriz psf* magnitudes. This data-set was shuffled to prevent unwanted correlations which might be introduced by a previous ordering of the objects. We ended up with 15 input features, by creating all possible color combinations and using the plain magnitudes, too. Even though, only the Euclidean distance based nearest neighbor model demands a renormalization of the input features, a min-max normalization was applied to all features of the whole data-set. The data-set was split in 30,000 objects for training and 50,000 objects for testing. As a performed k-fold cross validation did not show significant differences between the individual results, only the results of a single fold are presented.

4.2. *Redshift Estimation Methods*

The estimation of the photometric redshift distributions was done with three different methods that are described in more detailed, next. All models generate a mixture of Gaussian components to represent the *PDF*s of the redshift estimates.

4.2.1. *Nearest Neighbor Prediction Model*

For each object in the test sample, we have to generate an uncertain photometric redshift estimate. Similarly to the Gedankenexperiment in Section 3, the 256 nearest neighbors from the reference/training sample are extracted. Thereby the neighborhood is defined by the Euclidean distance within the 15-dimensional feature space. Instead of dealing with the mean value of the spectroscopic redshifts of the neighboring objects, a Gaussian mixture model with 5 components is fitted to the redshift distribution of the neighboring references. This representation is continuous and more smoother than dealing with an ensemble of individual redshift values. Furthermore, there are some computational advantages and it provides a compression of the amount of data that has to be stored and handled.

4.2.2. *Random Forest Prediction Model*

Similarly to the nearest neighbor approach, the random forest approach makes use of a partitioning of the high-dimensional feature space. 256 individual decision trees are build where bootstrapping and bagging ensure different feature space partitionings. Hereby the training data is used to generate the individual trees. For each of the test objects, the spectroscoping redshift of the final leaf of each decision tree is determined. Those reference redshift values are used to fit a Gaussian mixture model, likewise to the nearest neighbor approach. The resulting mixture model is considered to reflect the estimated redshift *PDF*. Both, the nearest neighbor and the random forest approach are local within the feature space and therefore do not provide a generalization. The next

https://doi.org/10.1017/S1743921316013089 Published online by Cambridge University Press

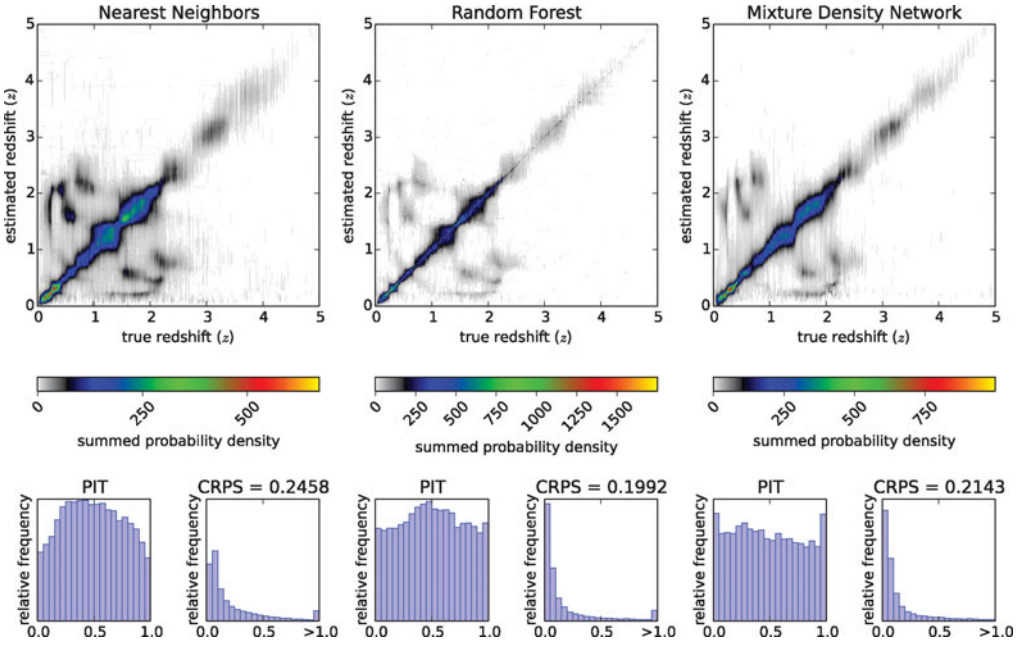

Figure 3. Comparison of the performance of Nearest Neighbors, Random Forest and Mixture Density Network based uncertain redshift estimates. The upper plots show a comparison between the estimated redshift and the true spectroscopic redshift. In contrast to plotting the point estimates against the true values, a probability distributy is plotted in y-direction. The lower plots visualize the *PIT* and *CRPS* values.

method is generalizing much better, but hence might not get all tiny difference that might be locally reflected by the reference data.

4.2.3. *Mixture Density Network Prediction Model*

A mixture density network is a modification of a simple multilayer perceptron. Instead of training to predict a single value, this kind of network predicts in the parameter of a mixture model. In our case we used a Gaussian mixture model, as it was used in both other approaches. Based on the means, sigmas and weights of the Gaussian component, the *CRPS* is used as the loss function for training. This ensures, that during training the weights of the network architecture are modified to minimize the *CRPS* and hence produce better results. By applying an early stopping technique, effects caused by overfitting are minimized.

4.3. *Results*

The results of the three different *PDF* generating approaches are presented in Figure 3. All three results show a very nice symmetry. This is a clear indication that multimodality is recovered in bot directions. E.g. some objects at $z = 0.5$ have a certain probability of being at $z = 1.5$ and vice versa. With respect to *CRPS*, the random forest based estimation shows the best performance. Taking into account the number of variables required to represent the reference objects / decision trees / weight and bias matrices, this is easy to understand. Both, the nearest neighbor and the random forest approach are slightly overdispersed. This indicates, that the estimated distributions are to broad with respect to their deviations from the true redshifts. The mixture density network instead generalizes much better and shows nearly a perfectly calibrated *PIT*.

https://doi.org/10.1017/S1743921316013089 Published online by Cambridge University Press

5. Conclusions

Multimodal Redshifts: We presented and motivated the necessity of multimodal redshift probability distributions. Especially in the case of estimating the redshift based on broadband photometry with objects that cover a larger redshift range and have errors in their measurements, multimodal probability distribution have to be expected. Those estimates can be either represented by an ensemble of individual estimates, by an approximation via a mixture model or by providing a full *PDF*. We have to learn how to propagate those uncertain estimates correctly, instead of using point estimates only.

Proper Tools and Measures: To have a correct and fair comparison between different approaches, we have to built reference data-sets and have clearly defined redshift estimation challenges. Further more, a standard set of proper scoring measures and tools to evaluate the performances of estimated probabilistic redshift distributions is demanded. We presented two well accepted proper tools from the field of weather forecasting that should be used in astronomy, too.

Future Work: In order to improve the quality of the predictions, we currently work on combining deep convolutional networks with mixture density networks. The work presented by Antonio D'Isanto at this meeting shows how well we can produce *PDF*s based on imaging data only. This enables us to skip the challenging steps of feature extraction and feature selection and makes a discrimination between quasars and galaxies in an extra step obsolete.

Acknowledgements

This work is based on data provided by the SDSS. SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration. Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is http://www.sdss3.org/.

References

Bishop, C. M. 2007, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1 edn. Springer

Dawid, A. P., 1984, *Journal of the Royal Statistical Society. Series A (General)*, 278,

Gieseke, F., Polsterer, K. L., & Zinn, P.-C. 2012, *Astronomical Data Analysis Software and Systems XXI*, 461, 537

Gieseke, F., Polsterer, K. L., Oancea, C. E. & Igel, C. 2014 *European Symposium on Artificial Neural Networks (ESANN)*, 87,

Gneiting, T., Raftery, A. E., Westveld III, A. H., & Goldman, T. 2005, *Monthly Weather Review*, 133, 1098

Gneiting, T., Balabdaoui, F., & Raftery, A. E. 2007 *Journal of the Royal Statistical Society: Series B*, 69, 243

Gneiting, T. & Raftery, A. E., 2007, *Journal of the American Statistical Association*, 102, 359

Grimit, E. P., Gneiting, T., Berrocal, V. J., & Johnson, N. A., 2006, *Quarterly Journal of the Royal Meteorological Society*, 132, 2925

Heinermann, J., Kramer, O., Polsterer, K. L., & Gieske, F. 2013 *Annual Conference on Artificial Intelligence*, 86,

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

Polsterer, K. L., Zinn, P.-C., & Gieseke, F. 2013, *MNRAS* , 428, 226

Polsterer, K. L., Gieseke, F., Igel, C., & Goto, T. 2014, *Astronomical Data Analysis Software and Systems XXIII*, 485, 425

Polsterer, K. L. D'Isanto, A., & Gieseke, F. 2016 *MNRAS*,

Schneider, D. P., Richards, G. T., Hall, P. B., *et al.* 2010, *The Astronomical Journal*, 139, 2360

Discussion

Giuseppe Longo: Comment: In the Euclid mission, photometric redshift estimation play an important role and therefore needs a proper treatment, as presented in the talk. Selecting objects

Kai Polsterer: Comment: For photometric redshift estimation challenges, we should define a set of accepted proper scoring rules to ensure a fair comparison of methods.

Ray Norris: Question: Should the surveys be preserving the *PDF*s and how should the values be stored and returned? Currently just single values are returned.

Kai Polsterer: Answer: We should start defining standards to preserve *PDF*s and have databases methods that allow to transform the *PDF*s into the format we need. Sometime the mean could be sufficient, sometimes we could make use of a mixture of Gaussians and sometimes the probability of being in a certain redshift range have to be returned.

Ray Norris: This will require a change in the way the community deals with photometric redshifts.

Kai Polsterer: Yes, indeed, and this is probably more complex than providing standards and interfaces.

Eric Feigelson: Question: When fitting Gaussian mixture models the number of components can be estimate via the Bayesian information criterion. Why did you pick a fixed number of Gaussians instead of using the most plausible one.

Kai Polsterer: Answer: We calculated both, the Bayesian information criterion and the Akaike information criterion when fitting the Gaussian mixture model. In order to keep the presentation simple, we choose to use a fixed number of components.

https://doi.org/10.1017/S1743921316013089 Published online by Cambridge University Press