Uncertain Photometric Redshifts

Kai Lars Polsterer, 1⋆ Antonio D’Isanto, 1 and Fabian Gieseke, 2

1 Heidelberg Institute for Theoretical Studies (HITS) gGmbH, Astrominformatics, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany
2 Radboud University, Institute for Computing and Information Sciences, Heyendaalseweg 135, 6525 AJ Nijmegen, the Netherlands

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT
Photometric redshifts play an important role as a measure of distance for various cosmological topics. Spectroscopic redshifts are only available for a very limited number of objects but can be used for creating statistical models. A broad variety of photometric catalogues provide uncertain low resolution spectral information for galaxies and quasars that can be used to infer a redshift. Many different techniques have been developed to produce those redshift estimates with increasing precision. Instead of providing a point estimate only, astronomers start to generate probabilistic density functions (PDFs) which should provide a characterisation of the uncertainties of the estimation. In this work we present two simple approaches on how to generate those PDFs. We use the example of generating the photometric redshift PDFs of quasars from SDSS (DR7) to validate our approaches and to compare them with point estimates. We do not aim for presenting a new best performing method, but we choose an intuitive approach that is based on well known machine learning algorithms. Furthermore we introduce proper tools for evaluating the performance of PDFs in the context of astronomy. The continuous ranked probability score (CRPS) and the probability integral transform (PIT) are well accepted in the weather forecasting community. Both tools reflect how well the PDFs reproduce the real values of the analysed objects. As we show, nearly all currently used measures in astronomy show severe weaknesses when used to evaluate PDFs.

Key words: methods: statistical, methods: data analysis, techniques: photometric, galaxies: distances and redshifts

1 INTRODUCTION
Determining the distance of an object has always been an important task in astronomy. This involves distances on all scales, from our solar system to distances on cosmological scale. Generating measures with high confidence requires dedicated instrumentation and an appropriate observation time. Unfortunately, obtaining precise spectroscopic distance measures for all objects is rendered impossible by both, the huge amount of required observation time and the necessary sensitivity of the instrumentation. Especially for cosmology, precise distances are required to understand e.g. gravitational lensing on cosmic scales (Hennawi & Spergel 2005; Scranton et al. 2005). For example, the success of the Euclid mission (Laureijs et al. 2010) will heavily depend on the availability of precise distance measures. The distance on cosmological scale, the so called redshift $z$ that is caused by the expansion of the universe, has therefore be determined through alternative methods.

Based on the huge amount of data that is made available by dedicated survey mission, statistical methods can be used to infer redshifts. Typically, a wide range of photometric measurements are extracted from the imaging surveys like SDSS (York et al. 2000), UKIDSS (Lawrence et al. 2007), and WISE (Wright et al. 2010). Due to the usually selected broadband filters, multiple measurements allow to very broadly reconstruct the spectral energy distribution (SED). In the past, SED fitting approaches (as Bolzonella et al. 2000) have provided very good estimations for the redshift based on photometric measurements. Csabai et al. (2003) utilized a variety of techniques to generate photometric redshift estimates for the early data release of SDSS, while Beck et al. (2010) provide redshifts for the latest release. Besides classical template fitting approaches, more and more machine learning approaches are used. Instead of using a fixed set of template spectra those approaches are based on models trained on a given reference set (Laurino et al. 2011; Polsterer et al. 2013; Brescia et al. 2014). There is a wide range of algorithms available including artificial neural networks, random forest, nearest neigh-
we will discuss Hastie et al. (2001) introductions to various machine learning techniques are provided.

Instead of just generating a point estimate for a photometric redshift, some approaches can generate a probability density function (PDF). In contrast to dealing with just a single prediction, this enables scientists to evaluate the likelihood for different redshifts. By integrating over a certain region of the PDF, the probability of being in a certain redshift range can be directly calculated. This measure of likelihood improves the analysis and further usage of the predictions, as the uncertainty can be quantified and propagated correctly. Important for a PDF is that the integral is always one, otherwise it would not be a correct density distribution.

The generation and evaluation of PDFs introduces new challenges. In this work we introduce proper tools to astronomy, that allow a fair comparison of the prediction quality achieved through PDFs. By discussing common errors, we want to emphasize the necessity of proper tools and measures. In addition, we present two simple approaches for generating PDFs based on well known algorithms, a nearest neighbours approach and a random forest regression model. We do not claim, that these methods produce the best results, they are just simple examples for generating PDFs. Feature selection and post-processing are appropriate tools to further improve the generated PDFs (Polsterer et al. 2014).

Outline: After this brief introduction to the topic of photometric redshift estimation, in Section 2 we will discuss proper tools for estimating the quality of PDFs. Next we present two straightforward approaches for creating PDFs based on well known algorithms (see Section 3). In Section 4 the presented approaches and evaluation tools are used to demonstrate their performance in some simple redshift estimation experiments. We also discuss the results of the experiments in detail. Finally, in Section 5, we conclude our work.

2 QUALITY ESTIMATION TOOLS

When it comes to evaluating PDFs, one has to rethink the tools used for measuring the quality of an estimation. For the task of photometric redshift estimation, we make use of reference data taken via spectroscopy. Therefore we have training, test and evaluation data-sets that map photometric features to highly certain and unique redshifts. This allows us to measure the performance of our redshift estimation models very precisely. To be able to compare our results we should use a common set of measures and tools.

One of the currently proposed extensions for generating PDFs, is to simply add an uncertainty interval to a generated point estimate. Assuming a normal distribution, one could use this formulation to calculate a certain likelihood for every possible redshift. The task of evaluating this distribution with respect to a single photometric redshift cannot be done by using the measures that have been used in the past. We propose a new set of measures and tools that are capable of evaluating PDFs with respect to a single true value. Before giving a detailed description we would like to point out some common mistakes with respect to the PDF evaluation process.

2.1 Common Mistakes

In the past a lot of measures have been used to quantify the characteristics of the error distribution. It was easy to measure the deviation between the true value and an estimated value. Normalization with respect to the redshift was commonly applied to express a relative error. The mean/median of the error distribution was a very simple statistical measure to check for biases, while other measures were used to quantify the spread across the deviations. Those measures are well suited for analysing and describing the deviations between point estimates and true values.

2.1.1 Oversimplifying PDFs

Most of these measures have been commonly applied and are well understood and accepted by the community. Therefore it can be understood that they are applied in modified versions to evaluate PDFs. In order to apply those well known measures, e.g. the mean of a PDF is extracted and used as a single value for comparison. A simple example of why this approach is not appropriate is shown in Figure 1. Multimodal distributions are common for photometric redshift estimation problems, due to the low photometric wavelength resolution, differences in the nature of the underlying physical sources, and measurement uncertainties. In Section 3, we present a straightforward approach to generate PDFs. This approach clearly demonstrates that we have to expect multimodal distributions. In case of a simple bimodal distribution it is clear that the mean is likely to be in a region in between of the two components and therefore will exhibit a relatively low probability density. Calculating the root mean square error between the mean of the PDF and the true value is hence producing useless information. A measure based on the mean would misleadingly indicate a good performance, as soon as a non symmetric PDF is evaluated. To prevent confusion, especially if the true nature of the produced PDF’s is unclear, we do not recommend the use of this measure, es-
especially when comparing the results of different approaches. As the PDFs are defining a probability distribution, this information must be used accordingly and not over-simplified because of a preferred measure.

2.1.2 Wrong Definition of Outliers

Catastrophic outliers with respect to the error are considered to be bad when it comes to estimating photometric redshifts. In the past, an outlier was defined to be a point estimate that is too far away from its true redshift. The distribution of the deviations between the estimated and the true redshift value allowed a simple characterisation of those outliers by defining simple thresholds. Typically a histogram was used to inspect the distribution of the deviations. In Laureijs et al. (2011), the requirements of the Euclid mission are collected. The normalised standard deviation of the photometric redshift estimation $\sigma_z/(1+z)$ is specified to be lower than 5 percent for no more than 10 percent of the objects. In this section, we do not want to discuss whether this target is reachable or not. We would like to highlight the problems that arise by sticking to such a measure of outliers, when using PDFs instead of point estimates. In Figure 2 we visualise the problems with such a measure. To be able to measure the amount of outliers based on a threshold criterion, we observe that typically PDFs are shifted by their true redshift and co-added. The stacked result of all evaluated PDFs allows to determine the amount of predictions that are exceeding a given range. This can be seen to be similar to the evaluation of the histograms created for the deviations of the point estimates. In our example we choose to have three different models that generate PDFs for two objects, respectively. Model One generates two narrow Gaussian shaped PDFs that are shifted against the true redshift. Even though the resulting stacked PDF shows no outliers, it is obvious that the result is terribly off, as the likelihood for the true redshift is practically zero. As soon as broader Gaussians are used, the true redshift is better represented as indicated by the higher likelihood at the true redshift. Even though both PDFs of Model Two are more likely, the stacked PDF shows already a certain area in the marked outlier region. In our last example we choose to stack two bimodal PDFs. As one can see from Model Three, the resulting stacked PDF has more than 50 percent of its area outside the given range. Even though, each PDF has a very high density at the true redshift, the stacked version would indicate a catastrophically high rate of outliers.

When dealing with PDFs, outliers should be defined based on probability and not on the area of the stacked PDF. Even though the similarity is very high to the method commonly applied to point estimates, it cannot be used to quantify outliers based on PDFs. A PDF that is excluding the true redshift by showing a too low likelihood at this point, must be considered a catastrophic outlier instead. Not the relative difference between the true redshift and parts of the PDF, but the likelihoods should be used as an outlier criterion.

2.2 Likelihood

Given a PDF as a redshift estimation and the true redshift value $z_t$, it is possible to evaluate the likelihood of the PDF at the given value. This likelihood specifies how well the given PDF represents the true redshift value. A low likelihood clearly indicates a unlikely redshift. In case the likelihood of an estimated PDF at the true redshift is very low, the PDF is not a good representation. Therefore the likelihood can be used to express how well a PDF represents the true redshift value. By maximizing the likelihood of the PDF, we can improve our model.

Based on the likelihood, we can efficiently detect those redshifts where the PDF fails to represent the true redshift. A histogram of the likelihoods of the PDFs at the corresponding true redshift value $z_t$ visualizes this distribution and allows to define a cutoff criterion. Every PDF below such a threshold, can be considered a catastrophic outlier with respect to the photometric estimation.

2.3 Probability Integral Transform

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 PDFs 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.

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 PDFs have to be transferred into continuous density functions (CDF)s (see Equation 1).

\[
CDF(z_t) = \int_{-\infty}^{z_t} PDF(z)dz
\]  

(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). Hereby $t \in \{1, 2, \ldots, N\}$ indexes the corresponding tuple of a predicted redshift distribution and the matching true redshift for $N$ data items.

\[
p_t = CDF_t(z_t)
\]  

(2)

In Appendix B an example is provided on how to generate a PIT histogram based on plain predictions. When the analysed PDFs are of Gaussian nature with $\mu$ and $\sigma^2$ as mean and variance, the CDFs can be evaluated by using Equation 3. For a Gaussian mixture model, the corresponding CDF is an additive mixture of single Gaussian CDFs multiplied with their weights, respectively.

\[
CDF(z_t) = \frac{1}{2} \left[ 1 + \text{erf} \left( \frac{z_t - \mu}{\sqrt{2}\sigma^2} \right) \right]
\]  

(3)

In case the predictions are ideal, the distribution of $p_t, t \in \{1, 2, \ldots, N\}$ has to be uniform. As shown in Figure 3, multiple aspects can be verified by plotting the histogram of this distribution. Only if the distribution of $p_t$ exhibits a uniform shape, the PDFs 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 PDFs that generate a peaked, concave histogram. As soon as a bias is present in the PDFs, a slope is added to the distribution of $p_t$.

The histogram of the PIT values allows to visually check the calibration and sharpness. It provides intuitive access to multiple aspects of the PDFs with respect to the corresponding true redshifts. Therefore we recommend this tool to be always used when evaluating PDFs.

2.4 Continuous Ranked Probability Score

When comparing the performances of different approaches that generated PDFs based on photometric features, a proper score should be used to measure the individual performances. Please see Gneiting & Raftery (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:

\[
CRPS = \frac{1}{N} \sum_{i=1}^{N} crps(CDF_i, z_i),
\]

with \( crps(CDF_i, z_i) = \int_{-\infty}^{\infty} [CDF_i(z) - CDF_i(z)]^2 dz \) (4)

where \( \Phi \) and \( \phi \) represent the CDF and PDF of a normal distribution with mean 0 and variance 1, respectively. In Equation 6 the \( \frac{\Phi(x)}{\sqrt{\pi}} \) 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 distributions. A Gaussian mixture model (see Equation 7) defines a distribution as a combination of \( M \) number of Gaussians with independent means \( \mu \) and variances \( \sigma^2 \).

\[
GMM(\mu, \sigma^2, \omega) = \sum_{i=1}^{M} \omega_i N(\mu_i, \sigma^2),
\]

with \( \sum_{i=1}^{M} \omega_i = 1 \) and \( \omega_i \geq 0 \) for \( i \in \{1, 2, \ldots, M\} \) (7)

whereby the weights \( \omega_i \) control the contributions of the individual Gaussians to the final distribution. With

\[
A(\mu, \sigma^2) = 2\sigma \phi(\frac{\mu}{\sigma}) + \mu \left[ 2\phi(\frac{\mu}{\sigma^2}) - 1 \right],
\]

we can calculate the CRPS of a GMM (Grunt et al. 2006). 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.

3 GENERATING UNCERTAIN REDSHIFTS

In the past, multiple approaches have been developed and applied to whole zoo of photometric features to generate redshift point estimates. Some of those approaches have been extended to generate PDFs by either sampling over the input uncertainties or adding fixed uncertainties to the point estimates. When fitting spectral energy distributions to the photometric features, under certain limitations, PDFs can be derived directly. Instead of comparing or discussing those approaches, we would like to introduce two simple but powerful approaches to generate PDFs.

3.1 Nearest Neighbours based PDF

To generate a point estimate, nearest neighbour based approaches have proven to generate good results. Based on the redshift of the nearest neighbour a point estimate is generated, usually by either calculating the mean or the weighted mean. This approach performs very well for a large set of reference objects, as a larger set is more likely to nicely represent all the objects in the high-dimensional feature space. Because this approach can be seen as sampling the feature space in the vicinity of an object, we assume that this sampling can be used to estimate the density of the distribution with respect to the redshift. Instead of calculating a single point estimate, we use the neighbours as a density representation.

By fitting a Gaussian mixture model (GMM) we can transfer the individual redshift values of the neighbours into a Gaussian representation. Therefore we can use the nearest neighbours to produce a PDF instead of a point estimate. See Figure 5 for an example of fitting a GMM to the nearest neighbours redshift distribution in order to generate a
PDF. This example is based on the task of fitting photometric redshifts of quasars taken form SDSS(DR7). When checking the distribution of the nearest neighbours, a multimodal distribution can be observed. In our example, a point estimate would be in the redshift region around $z = 1$ even though neither the true redshift nor the neighbours are close to $z = 1$. This is a clear indicator, why a plain point estimate will never be able to produce a good representation of the true redshift. An example on how to formalise this in Python is given in Appendix A.

It is obvious that this approach suffers from one limitation. In order to fit the GMM quite well, we need a large number of neighbours that are not too far away with respect to the applied distance measure in the feature space. Section 4 provides an analysis of the number of considered neighbours with respect to the number of fitted Gaussians. Especially those regions in the feature space that show a low number of reference objects because of physical or observational selection effects, will produce bad redshift estimations. One could consider weighting the individual neighbours when fitting the GMM, but still the number of neighbours that represent the redshift well, will be low. This will either smear out the resulting PDF by covering a too large range of references or produce inaccurate PDFs by overfitting the GMM. As this approach is very intuitive and helpful for comparing the performance of other approaches, we use it to retrieve a base performance measure. The approach presented next is not so sensitive to the discussed limitation.

### 3.2 Random Forest based PDF

A random forest is an ensemble of uncorrelated decision trees that are created based on a given set of reference objects. More details on random forest regression models can be found in Breiman (2001). Typically this ensemble is used to derive a mean of all members. Each decision tree is partitioning the feature space. Similarly to the nearest neighbour approach, the individual member values can be interpreted as a part of a redshift distribution. Therefore we fit a GMM directly to the results of the individual decision trees, instead of deriving just a mean value. In Appendix A we describe how to modify the nearest neighbour example to use a random forest regressor.

In contrast to the nearest neighbour approach, we are not strictly bound to the number of nearest neighbours to sample our high-dimensional feature space. The randomisation during the creation of the decision trees via feature selection, bootstrapping, and bagging creates independent trees which partition the feature space. By producing a larger number of trees, we are not smearing out the redshift information as we did for the nearest neighbour approach. We do not sample directly from our reference set, we use independent representations instead.

In the experiments section this difference in sampling from the feature space is directly reflected by the performance of the resulting PDFs.

### 4 EXPERIMENTS AND RESULTS

In order to evaluate the performance of the introduced approaches, we will focus on the use-case of photometric redshift estimation of quasars. To provide results that are easily comparable to other publications we choose to use the photometry data of SDSS(DR7) based on the quasar catalogue by Schneider et al. (2010). We randomly selected two subsets of 30,000 objects each. One subset was used for training, while the other was used for validation purposes only. As these experiments are just of conceptual nature, we did not apply cross-folding techniques as we would do, when producing a scientifically relevant catalogue.

#### 4.1 Effect of Number of Neighbours and Number of Gaussians

As discussed in the previous section, the number of nearest neighbours is a limiting factor when fitting a GMM. In Figure 6, we compare the effects of varying the number of neighbours and the number of Gaussian components. We used the same object as it is used in Figure 5. With an increasing number of Gaussian components, the resulting GMM over-fits. This is especially true, when the number of available nearest neighbours is too low. Single redshift values might be represented by a single Gaussian with an extremely low variance. Therefore the underlying redshift distribution is not represented well and should not be considered for scientific use. In the multimodal case, a too low number of Gaussian components is generalising too much. Two separated modes in the redshift distribution cannot be reconstructed by a single Gaussian. As one can see, the risk of over-fitting the GMM decreases with an increasing number of nearest neighbours. The down side of increasing the number of neighbours is the decrease in sharpness of the redshift distribution.

We found that the best results are obtained with a medium number of neighbours and a medium number of Gaussian components. This number differs as soon as the number of available reference objects and the coverage of the high-dimensional feature space changes.

Figure 5. PDF of a quasar at $z = 1.959$ calculated with a reference set of 30,000 high redshift quasars taken from SDSS(DR7). Based on the 64 nearest neighbours only, a histogram of the redshift distribution is generated (background). The individual redshifts of the neighbours are marked at the upper edge. The fitted Gaussian mixture model and its components are plotted in the foreground. Note the obvious multimodal distribution of the neighbours that reproduces the true redshift (marked in red) quite well.
Using just a single Gaussian does not allow to represent the underlying multimodal redshift distribution. As one can see, a too low number of Gaussian components results in a bad representation of the underlying PDF. In this example, 16, 64, and 1024 neighbours are used to fit 1, 2, 5, and 15 Gaussian components, respectively.

**Figure 6.** Comparison of resulting PDFs obtained by varying the number of considered neighbours and varying the number of fitted Gaussian components. In this example, 16, 64, and 1024 neighbours are used to fit 1, 2, 5, and 15 Gaussian components, respectively. Using just a single Gaussian does not allow to represent the underlying multimodal redshift distribution. As one can see, a too low number of considered neighbours with respect to the number of fitted Gaussian components results in a bad representation of the underlying redshift distribution. In contrast, too many neighbours smear out the redshift distribution and therefore lead to much broader PDFs that exhibit a much lower likelihood at the true redshift (marked red).

**Figure 7.** Comparison of different nearest neighbour based PDFs. The 64 nearest neighbours are used to create (from left to right): [i,ii] PDFs with a single Gaussian at the mean of the neighbours, with a fixed standard deviation of $\sigma = 0.01$ and $\sigma = 0.2$, respectively. [iii] A single Gaussian at the mean of the neighbours with an adaptive standard deviation based on the standard deviation of the neighbours. [iv] A GMM with five components, fitted to the 64 nearest neighbours. Instead of using the usual diagnostic diagram of plotting the true redshift against the estimated one, we plot a density distribution for each object at its true redshift. In addition, both the PIT and the CRPS are presented.
4.2 Nearest Neighbour PDFs

We evaluated the PDFs that have been generated by four different nearest neighbour based approaches. In our experiment we choose 64 nearest neighbours to be used to generate the PDFs. In Figure 7, the results of this comparison are presented. The different approaches are:

(i) PDFs, based on a single Gaussian with the mean of the nearest neighbours and a fixed, very small standard deviation of $\sigma = 0.01$. Because of the used extremely narrow single Gaussian, the results can be seen to mimic a single point estimate.

(ii) PDFs, based on a single Gaussian with the mean of the nearest neighbours and a fixed, broad standard deviation of $\sigma = 0.2$.

(iii) PDFs, based on a single Gaussian with the mean of the nearest neighbours and an adaptive standard deviation which is the standard deviation of the nearest neighbours.

(iv) PDFs, based on a GMM with five Gaussian components that has been fitted to the nearest neighbours.

In order to compare the results, we modified the usually applied diagnostic diagram in which the true redshift is plotted against the estimated one. Instead of using a single estimated redshift, we use the density distribution of the PDFs to plot an intensity parallel to the y-axis. The resulting plot, represents the average behaviour and the spread of all visualised PDFs. In addition, both, the PIT and the CRPS are presented in a histogram.

As one can see, the results of the single Gaussians with a fixed, very narrow standard deviation result in a diagnostic plot that is comparable to the ones from point estimates. The resulting PIT shows how extremely underdispersed the resulting PDFs are. The CRPS of 0.40 is the worst of all experiments.

When using the approach with the fixed standard deviation of $\sigma = 0.2$, the PIT still exhibits an underdispersion of the PDFs. Even though a quite high standard deviation was chosen, the CRPS of 0.34 indicates a better performance of the PDFs, which have been generated with this approach.

The PIT histogram of the approach with the adaptive standard deviation shows a slight overdispersion. This indicates that even with such a large reference set the number of neighbours is not sufficient to sample the redshift distribution. A CRPS value of 0.28 is a quite big improvement in comparison to the previous experiment. Due to the underlying physical model, some regions allow to specify the redshift more precisely, e.g. redshift regions where strong spectral features move from one band into another. Therefore the adaptive specification of the uncertainty is superior to a common uncertainty that is used for all objects.

By allowing for multimodalities, the results improve again. The PDFs that are based on a GMM with five components, show both, an improved PIT histogram and an improved CRPS value of 0.24. When inspecting the comparison between the estimated and the true redshifts, the accounting for multimodalities tremendously improves the density concentration towards the ideal diagonal of the plot. Hence it is important to mention, that such a plot is not a good tool to represent multimodal probability distributions. We just use this diagnostic plot, as it is very similar to the ones, commonly used for photometric point estimates.

| Table 1. Comparison of the CRPS values of all experiments. |
|------------------------------------------------------------|
| Nearest Neighbours | Random Forest |
| (CRPS) | (CRPS) |
| $N(\mu, \sigma = 0.01)$ | 0.3979 | 0.3118 |
| $N(\mu, \sigma = 0.2)$ | 0.3368 | 0.2711 |
| $N(\mu, \text{adaptive } \sigma)$ | 0.2816 | 0.2313 |
| $\text{GMM, } M = 5$ | 0.2415 | 0.1960 |

4.3 Random Forest PDFs

The experiments for measuring the performance of the random forest based PDFs was done with an ensemble of 256 decision trees. We choose four approaches that are similar to the ones used in Subsection 4.2:

(i) PDFs, based on a single Gaussian with the mean of the ensemble and a fixed, very small standard deviation of $\sigma = 0.01$.

(ii) PDFs, based on a single Gaussian with the mean of the ensemble and a fixed, broad standard deviation of $\sigma = 0.2$.

(iii) PDFs, based on a single Gaussian with the mean of the ensemble and an adaptive standard deviation which is the standard deviation of all trees.

(iv) PDFs, based on a GMM with five Gaussian components that has been fitted to the results of the individual trees. As discussed above, we could utilize more than 256 trees and therefore fit more Gaussian components, but we want to produce results that are still comparable to the nearest neighbours approach.

In Figure 8 the results of these four approaches are summarised. With the very narrow single Gaussian PDFs which mimic point estimates, the random forest approach already produces better results than the comparable nearest neighbours based approach. This is reflected by a better CRPS value of 0.31. The reason for this is that already 64 nearest neighbours are producing a mean estimate as they are covering a too large volume in the high-dimensional feature space. The PIT is very efficient in tracing the extreme underdispersion of the PDFs produced with the first approach.

The second approach with a fixed but broader standard deviation, produces already a CRPS of 0.27 that is comparable to the adaptive setting in the nearest neighbours experiment.

With an adaptive selection of the standard deviations for the single Gaussian PDFs, the multimodal approach of the nearest neighbours is already beaten with respect to the CRPS. For many objects, the nearest neighbours already cover a too large region in the redshift domain and therefore produce much broader PDFs than the random forest approach. The PIT histogram of this approach still shows an overdispersion, a clear indicator of a more complex underlying redshift distribution.

When using a GMM with five components, we are able to reconstruct the redshifts very well. The CRPS value went down to 0.20 in this experiment with a PIT histogram showing just a slight overdispersion. In Table 1 the resulting CRPS values of all experiments are summarized.
5 CONCLUSION

The problem with estimating photometric redshifts is the degeneracy of the redshift reconstruction problem, caused by the too low spectral resolution of the photometric features. We have shown, that in case of photometric redshift estimation, multimodal redshift distribution have to be expected. A probabilistic description of the redshift estimates allows the scientists to account for this and to propagate the uncertainties correctly.

With the proposed two approaches, PDFs can be generated in a straightforward way. Those PDFs can be used to describe photometric redshifts. As the presented approach is very general in can be applied to any kind of regression problem that should produce probabilistic estimates.

When dealing with PDFs, proper tools for evaluating the performance are required. We introduced the PIT and the CRPS as evaluation tools for photometric redshift estimation. In our experiments we could show, how well those tools help to understand and evaluate probabilistic redshift estimates. Uncertain photometric redshifts will play an important role very soon in many projects and experiments. Therefore we would recommend to make the presented tools a standard when comparing PDFs in astronomy.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support of the Klaus Tschira Foundation. The authors would like to thank Tilman Gneiting for the inspiring discussion on how to evaluate the quality of predictive distributions. The authors appreciate the fruitful discussion with Nikos Gianniotis on the importance of likelihoods. 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

Beck R., Dobos L., Budavári T., Szalay A. S., Csabai I., 2016, MNRAS, 460, 1371
Bishop C. M., 2007, Pattern Recognition and Machine Learning (Information Science and Statistics), 1 edn. Springer
Bolzonella M., Miralles J.-M., Pello R., 2000, A&A, 363, 476
Breiman L., 2001, Mach. Learn., 45, 5
Brescia M., Cavuoti S., Longo G., De Stefano V., 2014, A&A, 568, A126
Csabai I., et al., 2003, AJ, 125, 580
Dawid A. P., 1984, Journal of the Royal Statistical Society. Series A (General), pp 278–292
Gneiting T., Raftery A. E., 2007, Journal of the American Statistical Association, 102, 359
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
Grimit E. P., Gneiting T., Raftery A. E., 2006, Quarterly Journal of the Royal Meteorological Society, 132, 2925
Hastie T., Tibshirani R., Friedman J., 2001, the elements of statistical learning, Springer, New York
Hennawi J. F., Spiegel D. N., 2005, ApJ, 624, 59
Herschbach H., 2000, Weather and Forecasting, 15, 559
Laureijs R. J., Duvet L., Escudero Sanz I., Gondoin P., Lumb D. H., Oosterbroek T., Saavedra Criado G., 2010, in Space
APPENDIX A: SIMPLE PDF GENERATION EXAMPLE

In Section 3 we introduced two simple concepts on how to generate PDFs. Here we give more detailed information on how to implement those concepts in Python.

In the first lines of the code we make sure that all required packages are imported. The `numpy` package provides all matrix and vector manipulation functionalities required for scientific computing. To solve the regression task, the `sklearn.ensemble` package and the `sklearn.mixture.GMM` class are imported. From the `sklearn.spatial` package the k-dimensional tree structure is used to speed up the search for the nearest neighbours. Finally the `sklearn.mixture.GMM` class is required to express the individual results as a mixture of Gaussians.

```
import numpy
import scipy.spatial
from sklearn.ensemble import RandomForestRegressor
from sklearn.mixture import GMM

predictions = []
for i in range(len(predictions[:, 0])):
    myModel = GMM(n_components=5, min_covar=0.0001).fit(predictions[i])
    print(myModel.weights_
    print(myModel.means_
    print(myModel.covars_

    predictions.append(numpy.array(randomForest.estimators_[i].predict(testX)))

predictions = numpy.asarray(predictions).T
```

When replacing Lines 9-10 by the following lines, a random forest regressor instead of a nearest neighbours approach is used. First the random forest is created and the 100 decision trees are fitted to represent the given data. As each individual estimator of the random forest gives an individual result for each input, those results are retrieved individually and stored in a list. Finally this list of predictions has to be represented as a matrix and transposed to fit the required data-format.

```
    predictions = []
    for i in range(len(predictions[:, 0])):
        myModel = GMM(n_components=5, min_covar=0.0001).fit(predictions[i])
        myModel.means_
        myModel.weights_
        myModel.covars_

        predictions.append(numpy.array(randomForest.estimators_[i].predict(testX)))

    predictions = numpy.asarray(predictions).T
```

APPENDIX B: PIT EXAMPLE CODE

In Subsection 2, the PIT was introduced as a powerful tool to analyse the PDFs. This Python example shows how to easily create a PIT for a given set of predictions.

In the first two lines we import the packages used for scientific computing and for plotting. Both the `numpy` and the `matplotlib` package are often used in the astronomy community.

```
import numpy
from matplotlib import pyplot
```

Next, we create 10,000 synthetical objects with a random value $y$ between 0.0 and 6.0 sampled from an uniform distribution. As a prediction, we generate 100 samples from a normal distribution $N(\mu = 0, \sigma = 0.2)$ for each object, that can be seen as representatives of a PDF. Predictions were shifted with respect to the original values based on deviations that are sampled from a normal distribution with a sigma of 0.2. In the end we have 10,000 times 100 samples from Gaussians that are slightly deviating with their mean from the $y$ values.
n = 10000
y = numpy.random.uniform(low=0.0, high=6.0, size=(n))
error = numpy.random.normal(loc=0.0, scale=0.2, size=(n))
predictions = y[:, None] + error[:, None] +
               numpy.random.normal(scale=0.2, size=(n, 100))

To generate a PIT that is based on individual values is as simple as generating it based on a PDF. What is required is the cumulative probability at the value \( y \) that is compared with the PDF. To achieve this, the predictions are sorted per object. Next, the number of predictions below the corresponding \( y \) value is determined. With respect to the number of samples, the cumulative probability at the value \( y \) is calculated.

predictions = numpy.sort(predictions, axis=1)
PIT = numpy.zeros(n)
for i in range(len(predictions)):
PIT[i] = len(numpy.where(predictions[i] < y[i])[0]) *
        1.0 / 100.0

Finally the PIT is plotted by simply generating a histogram over all individual cumulative probabilities.

pyplot.figure()
pyplot.hist(PIT, bins=10)
pyplot.show()