Heidelberg Institute for Theoretical Studies
HITS
University of Copenhagen
Department of Computer Science

Probability Density Functions for Astronomy

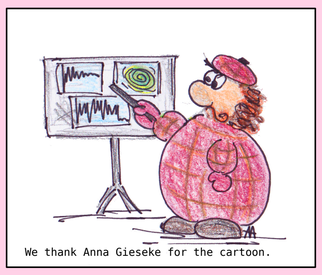
solving probabilistic regression problems with `ProbReg`

Kai Lars Polsterer[1]
Kai.Polsterer@h-its.org
Fabian Gieseke[2]
Fabian.Gieseke@di.ku.dk

[1]Heidelberg Institute for Theoretical Studies, Astroinformatics, Heidelberg, Germany [2]University of Copenhagen, Department of Computer Science, Copenhagen, Denmark

Abstract

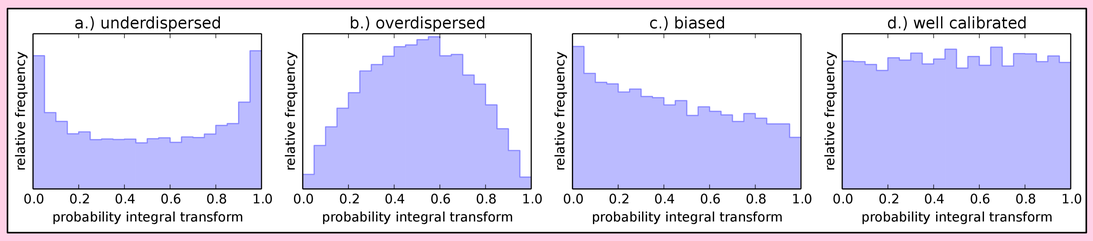
In many applications in astronomy, uncertainty quantification plays an important role. Probability density functions (PDFs) allow to quantify the likelihood of results and therefore enable scientist to produce better analysis results. We present a Python package to generate PDFs for regression tasks. Besides providing several functionalities to generate such PDFs, we present a whole tool set for evaluating the quality and visualizing the performance of the generated PDFs. Photometric redshifts are an important measure of distance for various cosmological topics. As spectroscopic redshifts are only available for a very limited set of objects, statistical regression models are helpful to derive estimates based on photometric measurements. We use the example of generating the photometric redshift PDFs of quasars from SDSS(DR7) based on psf- and model-magnitudes to present the functionalities of `ProbReg`.



Probability Integral Transform

$$CDF(z_{true}) = \int_{-\infty}^{z_{true}} PDF(z)\,dz$$

In 1984, the probability integral transform (PIT) was proposed by Dawid to be used to check the calibration and the sharpness of a predictive distributions. The PIT is a simple visual tool which is based on the cumulative distribution functions (CDF) at the true value. With respect to photometric redshift estimations, the PIT is plotted with the CDF of the estimated redshift at the true redshift z_{true}.



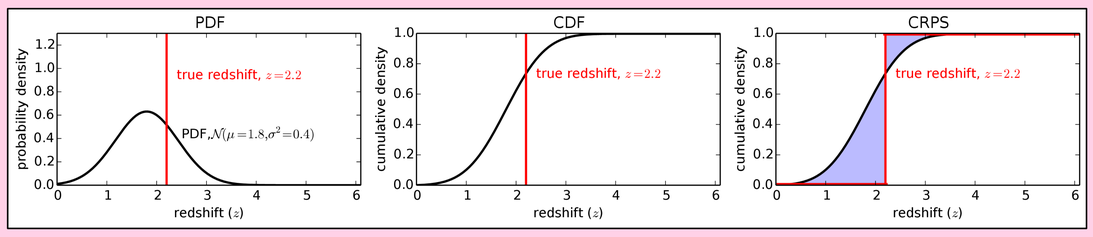
Example of four different probability integral transforms. In the case of under-dispersed PDFs a concave, u-shaped distribution is observed (a). A peaked, convex distribution is exhibiting over-dispersed PDFs (b). A slope in the PIT indicates that the analysed PDFs are biased (c). Only when the PIT exhibits a flat, uniform distribution, the PDFs are well calibrated (d).

Continuous Ranked Probability Score

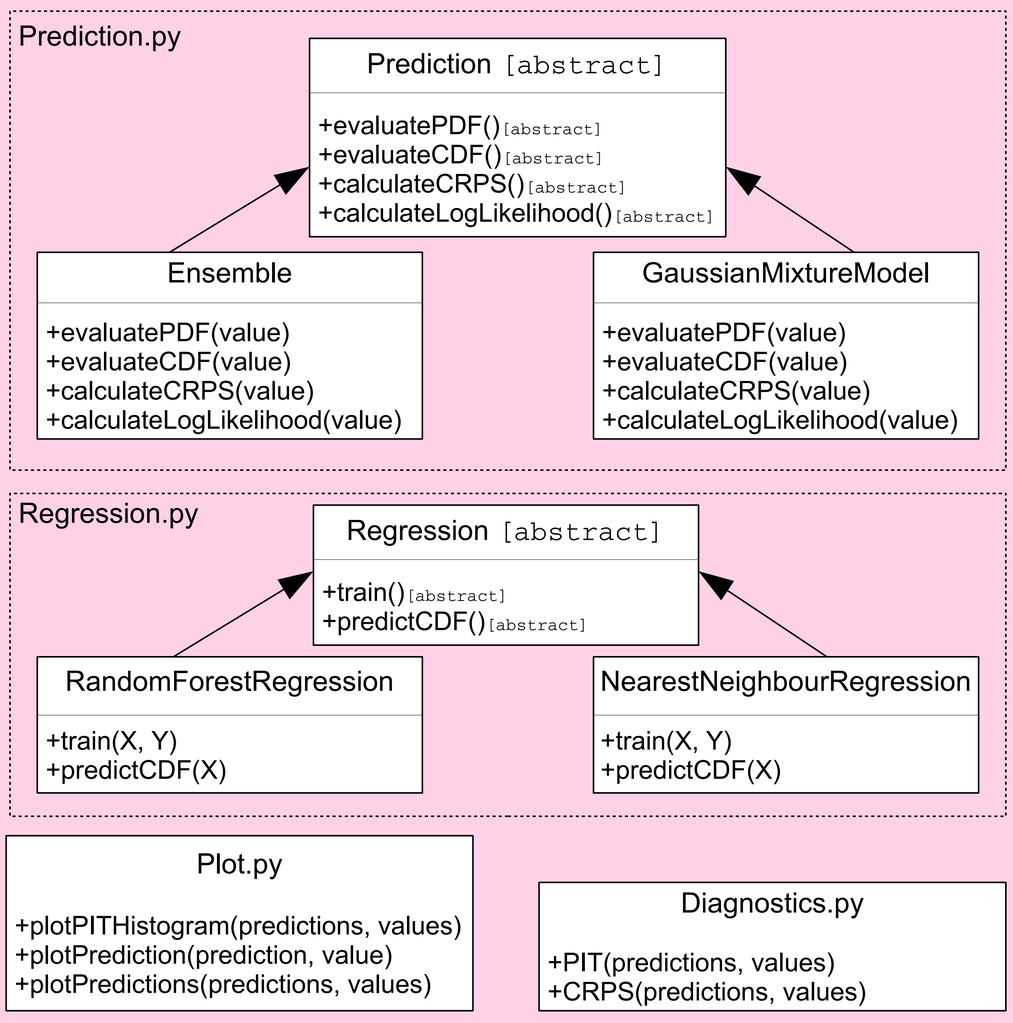
$$CRPS = \frac{1}{N}\sum_{i=1}^{N} crps(CDF_i, z_{true}),$$

$$with\ crps(CDF_i, z_{true}) = \int_{-\infty}^{+\infty} [CDF_i(z) - H(z_{true}, z)]^2\,dz$$

The continuous ranked probability score (CRPS) (Hersbach, 2000) is widely used in the field of weather forecasting for expressing the distance between the PDF and the true value. It can be used to compare a distribution with a single value as defined in the equation above. An illustration of the meaning of PDF, CDF, and CRPS is given below. The true redshift is plotted as a reference in red. The integral between the CDF and the Heaviside step-function H of the true redshift is the basis of the CRPS.

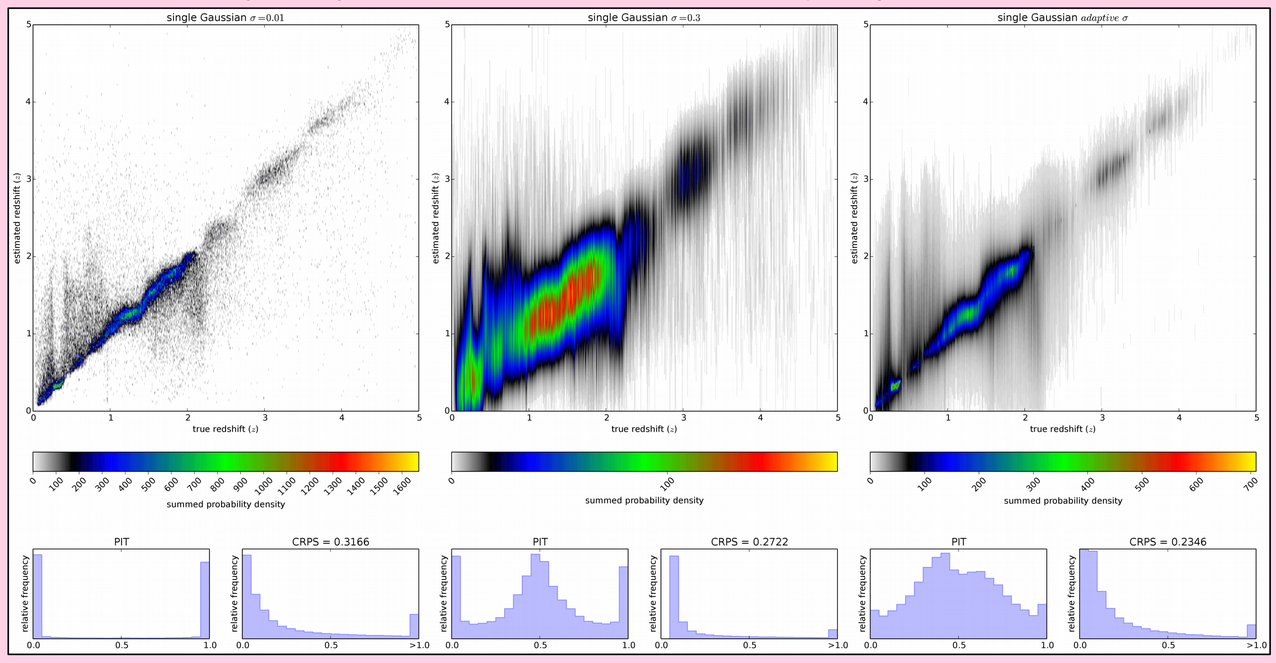


Class Diagram of `ProbReg`



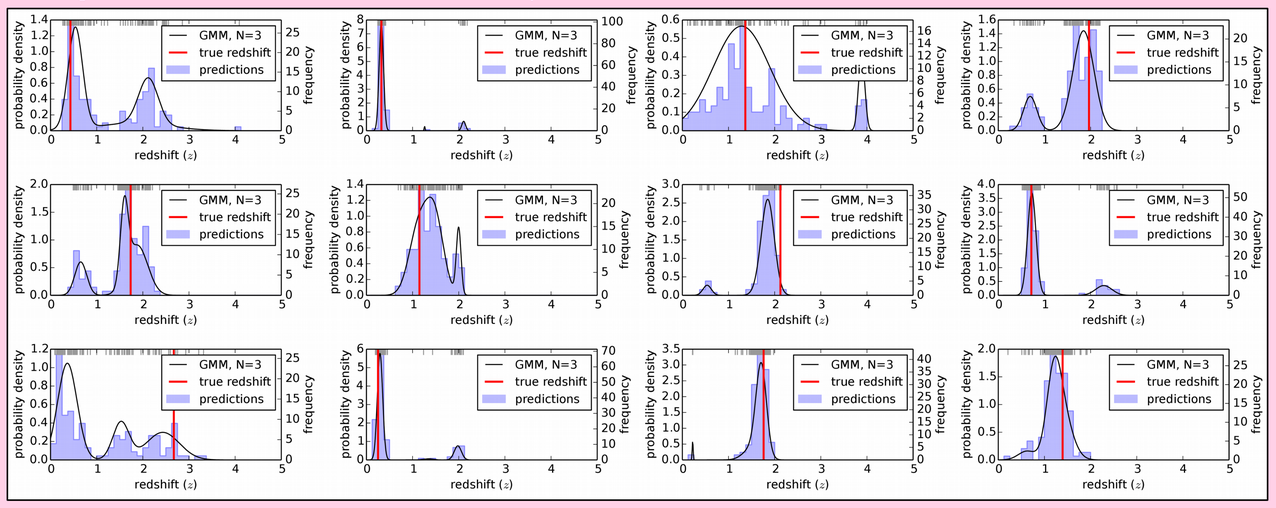
Presenting Uncertain Estimates

PDFs allow to capture the uncertainty of the estimates. Therefore the usual diagnostic plot - comparing the true and the estimated redshift - has to be changed to fit those requirements. We plot a density distribution for each object at its true redshift. In addition, both the PIT and the CRPS are presented. With a reference set of 30,000 high redshift quasar a PDF for all remaining 50,000 quasars from Schneider et al. (2010) has been calculated using `ProbReg`. The results are plotted through the provided plotting functionalities. Note that just a single Gaussian is used with either a fixed or an adaptive sigma values.

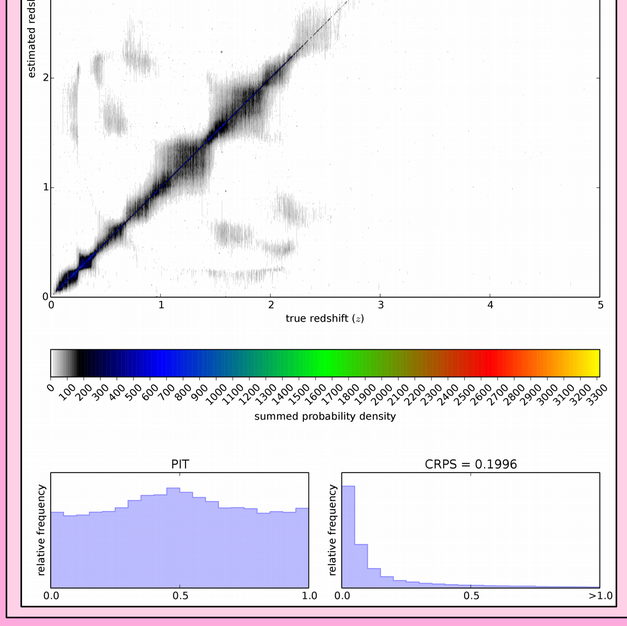


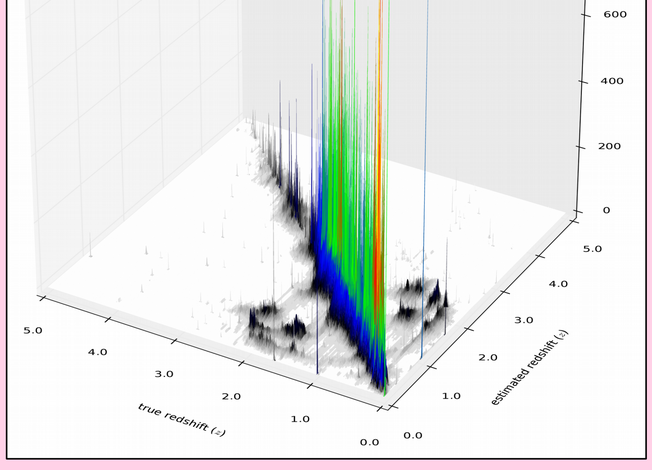
Multi-Modal Probability Density Functions

When estimating the photometric redshift based on a limited set of broad-band filters, multimodal PDFs have to be expected. This degeneracy is caused by the not sufficient observation of the actual spectral energy distribution and the noise of the individual measurements. Those challenges with multimodal PDFs have already been identified. In contrast to using a single estimate, PDFs enable scientists to evaluate the likelihood for every redshift and therefore propagate the uncertainties correctly. The individual PDFs of 12 quasars that have been calculated with a reference set of 30,000 high redshift quasars (SDSS DR7) are presented below. Based on the 128 nearest neighbours, the background histogram of the neighbour redshifts is generated with the individual values being shown at the upper edge. The fitted mixture model is composed of 3 Gaussians. An obvious multimodal distribution of the neighbours can be observed for some of the examples. As a reference, the true redshift is marked in red.



From the plot above, it is obvious that photometric redshifts require a multimodal probability representation. By using the multimodal capabilities of `ProbReg`, we can improve the results. The PDFs that are based on a Gaussian mixture model with five components, show both, an improved PIT histogram and an improved CRPS value (below). When inspecting the comparison between the estimated and the true redshifts, the accounting for multimodalities tremendously improves the density concentration towards the ideal diagonal and shows a nearly symmetric plot.





Kai Lars Polsterer

www.h-its.org/english/research/ain/

@AstroInformatix

ADASS 2016, Trieste, Italy, October 16 – 20, 2016