Calibrated Predictive Distributions for Photometric Redshifts

Biprateep Dey [1] **David Zhao** [2] **Jeffrey A. Newman** [1] **Brett H. Andrews** [1] **Rafael Izbicki** [3] **Ann B. Lee** [2]

Abstract

Many astrophysical analyses depend on estimates of redshifts (a proxy for distance) determined from photometric (i.e., imaging) data alone. Inaccurate estimates of photometric redshift uncertainties can result in large systematic errors. However, probability distribution outputs from many photometric redshift methods do not follow the frequentist definition of a Probability Density Function (PDF) for redshift — i.e., the fraction of times the true redshift falls between two limits z_1 and z_2 should be equal to the integral of the PDF between these limits. Previous works have used the global distribution of Probability Integral Transform (PIT) values to re-calibrate PDFs, but offsetting inaccuracies in different regions of feature space can conspire to limit the efficacy of the method. We leverage a recently developed regression technique that characterizes the local PIT distribution at any location in feature space to perform a local re-calibration of photometric redshift PDFs resulting in calibrated predictive distributions. Though we focus on an example from astrophysics, our method can produce predictive distributions which are calibrated at all locations in feature space for any use case.

1. Introduction

Galaxy distance, as measured by redshift, is essential for estimating intrinsic luminosity and 3D location in space, which is crucial information for many astrophysical studies. High-precision redshifts require resource-intensive observations and will only be feasible for a few percent of galaxies in upcoming photometric surveys. Thus, photometric redshifts (photo-z's)—redshifts estimated from imaging alone—will be necessary. Furthermore, accurate photo-z's are critical for some science cases (e.g., weak lensing cosmology), but PDFs from both main methods of photo-z estimation (galaxy spectral template-based and machine learning-based) fail to satisfy the frequentist definition of a PDF for redshift (Dahlen et al., 2013; Kodra, 2019; Schmidt et al., 2020). The fraction of times the true redshift falls between two limits z_1 and z_2 should equal the integral of a properly-defined PDF between these limits, for any arbitrary subset of the test data.

Current metrics used to measure the quality of calibration, like the distribution of the values of the cumulative distribution function (CDF) evaluated at the true redshift of the object (the Probability Integral Transform or PIT; see Eq. 1) can favor pathological but un-informative PDFs (Schmidt et al., 2020). Moreover, overall uniformity of PIT values is possible even if particular subsets of the same test data are poorly-calibrated (Zhao et al., 2021). If the PDFs are well-calibrated, then the distribution of the PIT values of a test sample will be uniform between 0 and 1 or their corresponding CDF will follow the identity line for any arbitrary subset of the test data. The same can be visualized with a P-P plot that shows the empirically calculated CDF versus their theoretical expected values. Ideally, the P-P plot should closely follow the identity line, but it often does not.

Several previous works have studied PDF re-calibration (e.g., Niculescu-Mizil & Caruana 2005; Rau et al. 2015; Kuleshov et al. 2018), though none can ensure that PDFs are well-calibrated at every point in feature space. Bordoloi et al. (2010) described a method to re-calibrate PDFs using a single correction factor based on the overall distribution of PIT values, which ensures a uniform global distribution of PIT values, but this single correction factor is applied to all PDFs and does not account for local variations. Importantly, these local inconsistencies in feature space can be detected using tests like the ones proposed in Jitkrittum et al. (2020) and Zhao et al. (2021), which we will leverage in our method.

In this work, we develop a local PDF re-calibration procedure that uses an estimate of the local distribution of PIT values (from Zhao et al. (2021)) to calculate a correction factor at any location in feature space. To demonstrate our method, we use the simulated data from Schmidt et al.

[1]Dept. of Physics and Astronomy and PITT-PACC, University of Pittsburgh, Pittsburgh, Pennsylvania, 15260
[2]Dept. of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, 15213
[3]Dept. of Statistics, Federal University of São Carlos (UFSCar), São Carlos, Brazil. Correspondence to: Biprateep Dey <biprateep@pitt.edu>.

ICML 2022 Workshop on Machine Learning for Astrophysics, Baltimore, Maryland, USA, 2022.

(2020), which has been used to compare photo-z CDE prediction methods in the past. We start with the marginal distribution of redshifts in the training set as our initial CDE estimate for all galaxies. Schmidt et al. (2020) demonstrated that such a CDE estimate can perform well on many commonly used metrics that check for calibration, although it does not provide information about individual galaxies and represents the worst case scenario. We use the "training set" from Schmidt et al. (2020) with about 44,000 objects as our calibration set; then split the remaining data into two sets:a validation set (twice as large as the calibration set) and a larger test set comprised of roughly 250,000 instances. We use one magnitude and 5-colors and their associated measurement uncertainties as the features. As the "true CDEs" are not known, we shall use data driven methods using the validation and test sets to evaluate the performance of our methods.

2. Re-calibration Procedure

Let $\widehat{p}(z|\mathbf{x})$ be the initial estimate of the true PDF $p(z|\mathbf{x})$ of the target variable z (redshift) given the input features $\mathbf{x}$ (galaxy colors and magnitudes). The random variable corresponding to z is denoted by Z. We define the local Probability Integral Transform (PIT) corresponding to this initial estimate as:

$$\widehat{\mathrm{PIT}}(z, \mathbf{x}) = \int_0^z \widehat{p}(z'|\mathbf{x})dz' = \widehat{F}(z|\mathbf{x}) \tag{1}$$

where $\widehat{F}$ is the cumulative distribution function associated with $\widehat{p}$. Using a labeled calibration set and a suitable regression method (monotonic neural networks; Wehenkel & Louppe 2019 in our case), we estimate the CDF of PIT values as a function of $\mathbf{x}$ following the method described in Zhao et al. (2021). The regression algorithm takes both the features and a value of coverage level (α; drawn from a $Uniform(0, 1)$ distribution while training) as inputs to get the CDF of PIT values ($r^{\widehat{p}}_{\alpha}(\mathbf{x})$):

$$r^{\widehat{p}}_{\alpha}(\mathbf{x}) \coloneqq P\left(\widehat{\mathrm{PIT}}(Z, \mathbf{x}) \leq \alpha|\mathbf{x}\right) = P\left(Z \leq \widehat{F}^{-1}(\alpha|\mathbf{x})|\mathbf{x}\right) \tag{2}$$

If our initial PDFs are locally calibrated, then the relation $r^{\widehat{p}}_{\alpha} = \alpha$ should hold for any $\mathbf{x}$, i.e., a plot of $r^{\widehat{p}}_{\alpha}$ vs. α (also called Amortized Local P-P plots or ALP plots) should closely follow the identity line. Most photo-z estimators do not produce locally calibrated PDFs and this relation does not often hold.

To re-calibrate the original PDF estimates $\widehat{p}(z|\mathbf{x})$ such that the relation $r^{\tilde{p}}_{\alpha} \approx \alpha$ holds for the new PDFs, $\tilde{p}$, for a new unseen test data set, we define, $\beta \coloneqq r^{\widehat{p}}_{\alpha}$ for each $\alpha \in \mathbb{G}$ and define a new cumulative distribution function, $\tilde{F}$, such that:

$$\tilde{F}^{-1}\left(\beta|\mathbf{x}\right) = \widehat{F}^{-1}\left(\alpha|\mathbf{x}\right) \tag{3}$$

Then by construction the new PDFs, $\tilde{p}$, will be calibrated since

$$r^{\tilde{p}}_{\beta}(\mathbf{x}) = P\left(Z \leq \tilde{F}^{-1}(\beta|\mathbf{x})|\mathbf{x}\right) = r^{\widehat{p}}_{\alpha} = \beta \tag{4}$$

Now, for $\tilde{z} = \tilde{F}^{-1}\left(\beta|\mathbf{x}\right)$ we will have,

$$\int_0^{\tilde{z}} \tilde{p}(z'|\mathbf{x})dz' = \beta \tag{5}$$

$$\implies \tilde{p}(\tilde{z}|\mathbf{x}) - \tilde{p}(0|\mathbf{x}) = \frac{d\beta}{dz'} = \frac{d\beta}{d\alpha}.\frac{d\alpha}{dz'} \tag{6}$$

Eqs. 3 and 1 imply

$$\tilde{p}(\tilde{z}|\mathbf{x}) = \tilde{p}(z|\mathbf{x})$$

and Eq. 2 implies

$$\frac{d\alpha}{dz'} = \widehat{p}(z|\mathbf{x})$$

It is not physical to have any object at redshift 0 so we can assume $\tilde{p}(0|\mathbf{x}) = 0$. This gives us the relation:

$$\tilde{p}(z|\mathbf{x}) = \widehat{p}(z|\mathbf{x}).\frac{d\beta(\alpha)}{d\alpha} \tag{7}$$

This means that our corrected PDF equals the initial PDF multiplied by a correction factor which is the local PIT distribution evaluated at the coverage corresponding to various redshifts. This relation is very similar to what Bordoloi et al. (2010) uses to re-calibrate photo-z PDFs except now the correction factor is calculated using the local PIT distribution rather than the empirical distribution obtained from the calibration set as a whole.

To numerically evaluate the correction factor, we use a spline based interpolator to calculate and evaluate the derivative of the output of the neural network evaluated on a dense grid of α for a given galaxy. The new re-calibrated PDFs are then normalized.

3. Results and Discussion

We apply our method to re-calibrate the initial CDE by learning the local distribution of PIT values by training $r^{\widehat{p}}$ on the calibration set and use it to re-calibrate the CDEs in our validation and test sets. To assess the quality of our re-calibrated CDEs, we train another regression model using the validation set and its re-calibrated CDEs. We infer the local CDF of PIT for every object in the test set before and after re-calibration using the two trained models. Fig. 1 (top) shows the diagnostic local P-P plot for five galaxies in the test set. The local P-P plots for the original CDEs diverge significantly from the identity line, showing the inadequacy of the original PDF. Global P-P plots for the

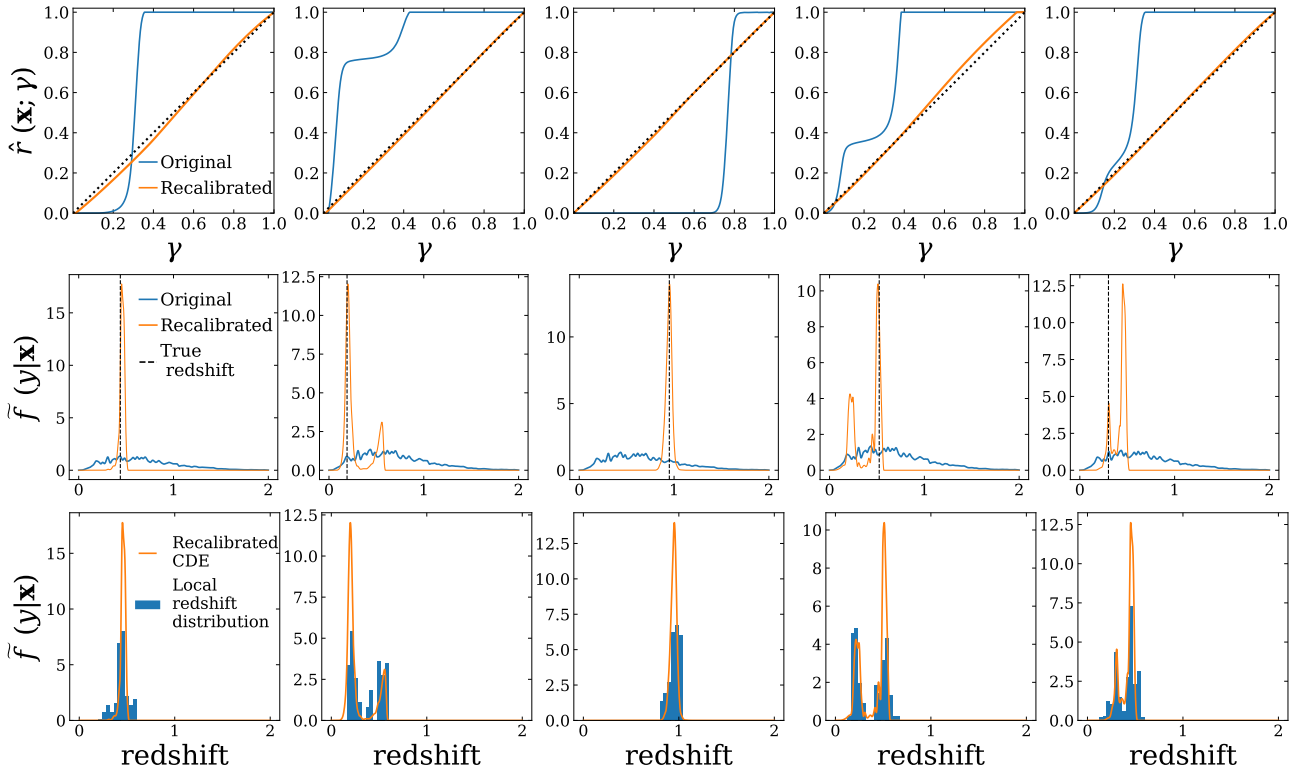

Figure 1. Top: Diagnostic local P-P plot for 5 galaxies before and after our re-calibration method is applied. The local P-P plots for the original CDEs diverge significantly from the identity line, showing the inadequacy of the original PDF. *Center:* CDEs for the corresponding galaxies before and after re-calibration along with their true redshift. Our method can recover multimodalities while ensuring good local calibration. *Bottom:* Rudimentary comparison of re-calibrated CDEs with the distribution of true redshifts of other galaxies having similar photometry. We show an inverse-distance weighted histogram of the redshift distribution of galaxies with similar photometry along with their CDEs. We observe that the histograms show bimodal distributions when our inferred CDEs are bimodal and unimodal when the inferred distribution is unimodal, matching expectations.

pathological initial CDE that we chose, would follow the identity line closely by construction (Schmidt et al., 2020). This highlights the importance of local tests of calibration. After re-calibration, the local CDF of PIT for these objects follow the identity line closely (i.e., the CDF of a uniform distribution), indicating good local calibration. Fig. 1 (center) shows the CDEs for the same galaxies before and after re-calibration. We not only observe that the re-calibrated CDEs are more informative but also multimodal CDEs can be recovered (as are typical for photo-z's), even when the input CDE before calibration is unimodal.

Due to noisy and limited information about redshift contained in galaxy images, galaxies with similar photometry may have different redshifts and vice versa. We want this property to be captured in photo-z CDEs, requiring them to be multimodal. As we do not know the "ground truth" CDEs, we generally have to rely on indirect methods to assess calibration. In Fig. 1 (bottom), we provide a rudimentary but direct demonstration that the CDEs we predict are indeed meaningful. We compare the CDEs of the five galaxies shown with the distribution of true redshifts of other galaxies with similar imaging data. We identify those counterparts by searching for other galaxies in the training set whose colors and magnitudes (rescaled by subtracting the mean and dividing by the standard deviation for each feature) lie within an Euclidean distance of 0.5 units of our selected galaxies. Fig. 1 (bottom) shows their redshift distribution as an inverse-distance weighted histogram. We observe that the histograms show bimodal distributions when our inferred CDEs are bimodal and unimodal when the inferred distribution is unimodal, matching expectations.

The Cramér-von Mises statistic between the local PIT CDF of each galaxy in the test set and the uniform distribution is a measure of the quality of conditional coverage (Schmidt et al., 2020), and decreases significantly on the entire test set after re-calibration (Fig. 2), with a mean decrease of $\sim 4.5\times$. We also see a large improvement in the value of the CDE Loss (Izbicki et al., 2017), which provides another

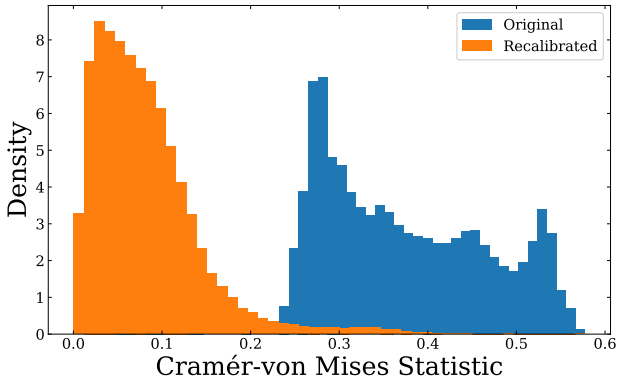

Figure 2. Distribution of the Cramér-von Mises (CvM) Statistic (i.e., mean squared difference) between the local PIT CDF of each galaxy in the test set and the CDF of a Uniform distribution. As the "ground truth" CDEs are unknown, we assess conditional coverage by training regression models to predict the local PIT CDFs on the calibration and validation sets. We observe a significant decrease in the value of CvM statistic for the entire test set, with the average value decreasing by $\sim 4.5\times$.

independent metric of conditional coverage, with a decrease from -0.84 to -10.71 after re-calibration. For comparison, in Schmidt et al. (2020) the photo-z algorithms considered yielded CDE losses ranging from -1.66 at worst to -10.60 at best.

This work shows that PDFs can be re-calibrated using local information and produce better uncertainty estimates. We show that our method works well even when the initial CDEs are uninformative. Our method might yield even better results if applied on initial CDE estimates from a reasonably good photo-z algorithm. This will be studied in a future work.

References

Bordoloi, R., Lilly, S. J., and Amara, A. Photo-z performance for precision cosmology. , 406(2):881–895, August 2010. doi: 10.1111/j.1365-2966.2010.16765.x.

Dahlen, T., Mobasher, B., Faber, S. M., Ferguson, H. C., Barro, G., Finkelstein, S. L., Finlator, K., Fontana, A., Gruetzbauch, R., Johnson, S., Pforr, J., Salvato, M., Wiklind, T., Wuyts, S., Acquaviva, V., Dickinson, M. E., Guo, Y., Huang, J., Huang, K.-H., Newman, J. A., Bell, E. F., Conselice, C. J., Galametz, A., Gawiser, E., Giavalisco, M., Grogin, N. A., Hathi, N., Kocevski, D., Koekemoer, A. M., Koo, D. C., Lee, K.-S., McGrath, E. J., Papovich, C., Peth, M., Ryan, R., Somerville, R., Weiner, B., and Wilson, G. A Critical Assessment of Photometric Redshift Methods: A CANDELS Investigation. , 775(2):93, October 2013. doi: 10.1088/0004-637X/775/2/93.

Izbicki, R., Lee, A. B., et al. Converting high-dimensional regression to high-dimensional conditional density estimation. *Electronic Journal of Statistics*, 11(2):2800–2831, 2017.

Jitkrittum, W., Kanagawa, H., and Schölkopf, B. Testing goodness of fit of conditional density models with kernels. In Adams, R. P. and Gogate, V. (eds.), *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, virtual online, August 3-6, 2020*, volume 124 of *Proceedings of Machine Learning Research*, pp. 221–230. AUAI Press, 2020. URL http://proceedings.mlr.press/v124/jitkrittum20a.html.

Kodra, D. The galaxy morphology-density relation at high redshift with candels. PhD Thesis, January 2019. URL http://d-scholarship.pitt.edu/35716/.

Kuleshov, V., Fenner, N., and Ermon, S. Accurate uncertainties for deep learning using calibrated regression. In Dy, J. G. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2801–2809. PMLR, 2018. URL http://proceedings.mlr.press/v80/kuleshov18a.html.

Niculescu-Mizil, A. and Caruana, R. Predicting good probabilities with supervised learning. In Raedt, L. D. and Wrobel, S. (eds.), *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*, volume 119 of *ACM International Conference Proceeding Series*, pp. 625–632. ACM, 2005. doi: 10.1145/1102351.1102430. URL https://doi.org/10.1145/1102351.1102430.

Rau, M. M., Seitz, S., Brimioulle, F., Frank, E., Friedrich, O., Gruen, D., and Hoyle, B. Accurate photometric redshift probability density estimation - method comparison and application. , 452(4):3710–3725, October 2015. doi: 10.1093/mnras/stv1567.

Schmidt, S. J., Malz, A. I., Soo, J. Y. H., Almosallam, I. A., Brescia, M., Cavuoti, S., Cohen-Tanugi, J., Connolly, A. J., DeRose, J., Freeman, P. E., Graham, M. L., Iyer, K. G., Jarvis, M. J., Kalmbach, J. B., Kovacs, E., Lee, A. B., Longo, G., Morrison, C. B., Newman, J. A., Nourbakhsh, E., Nuss, E., Pospisil, T., Tranin, H., Wechsler, R. H., Zhou, R., Izbicki, R., and LSST Dark Energy Science Collaboration. Evaluation of probabilistic photometric redshift estimation approaches for The Rubin Observatory Legacy Survey of Space and Time (LSST). , 499(2):1587–1606, December 2020. doi: 10.1093/mnras/staa2799.

Wehenkel, A. and Louppe, G. Unconstrained monotonic neural networks. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 1543–1553, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/2a084e55c87b1ebcdaad1f62fdbbac8e-Abstract.html.

Zhao, D., Dalmasso, N., Izbicki, R., and Lee, A. B. Diagnostics for conditional density models and bayesian inference algorithms. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 125, 2021.