TERM PROJECT FINAL REPORT

TROY RAEN

Abstract. Light reaching our telescopes from distant galaxies has been 'redshifted' by the expansion of space, and the magnitude of this shift increases with the distance to the galaxy. Accurate calculations of the quantities fundamental to cosmology rely on our ability to calculate this redshift for a large number of galaxies, using very low resolution data. Machine learning algorithms have shown success with this problem and are becoming the dominant solution method. In this work I study how the errors in redshift estimates scale with the size of the training set for two neural net architectures, a random forest algorithm, and a publicly available code developed specifically for photo-z's called GPz [1]. My code is available at `https://github.com/troyraen/photoz_errors`.

1. Introduction

It is well established that the light reaching our telescopes from distant galaxies is shifted toward the red end of the spectrum (relative to the frequency it was originally emitted at), and that the magnitude of this 'redshift' (denoted by z) increases with the galaxy's distance from us (e.g. see [6], [3], [5]). The combined measurements from many galaxies indicate that the universe itself, the space between galaxies, is expanding at a rate that increases with time. More precise calculations of this expansion rate, and several other quantities fundamental to cosmology, are being pursued, and they all depend strongly on our ability to make accurate redshift calculations for large numbers of galaxies.

The dataset used in this work is simulated and intended to mimic the data anticipated from the upcoming Large Synoptic Survey Telescope (LSST). LSST will collect data from large volumes of the sky and at rates several orders of magnitude above any other telescope to date. The community is making large efforts towards dealing with data at this scale, and one of these efforts is toward quick and accurate redshift calculations using machine learning techniques.

1.1. Dataset.

I use the dataset `Catalog_Graham+2018_10YearPhot` (see [3]) which consists of simulated telescope data: measurements of light in 6 frequency ranges (bins), plus errors on the measurements, for 3.8×10^6 galaxies. The dataset includes the true redshift for each galaxy, so this is a supervised, regression problem.

Figure 1 shows the true redshift distribution of the galaxies in the dataset. Projections along principal components that minimize the covariance in the data is shown in the table below. Figure 2 shows pairwise scatter plots of the features and the target along with correlation coefficients. I chose the two features with the highest absolute value correlations with the target redshift and show their scatter plot, colored by redshift, in Figure 3 to get a sense of the distribution.

PCA vectors

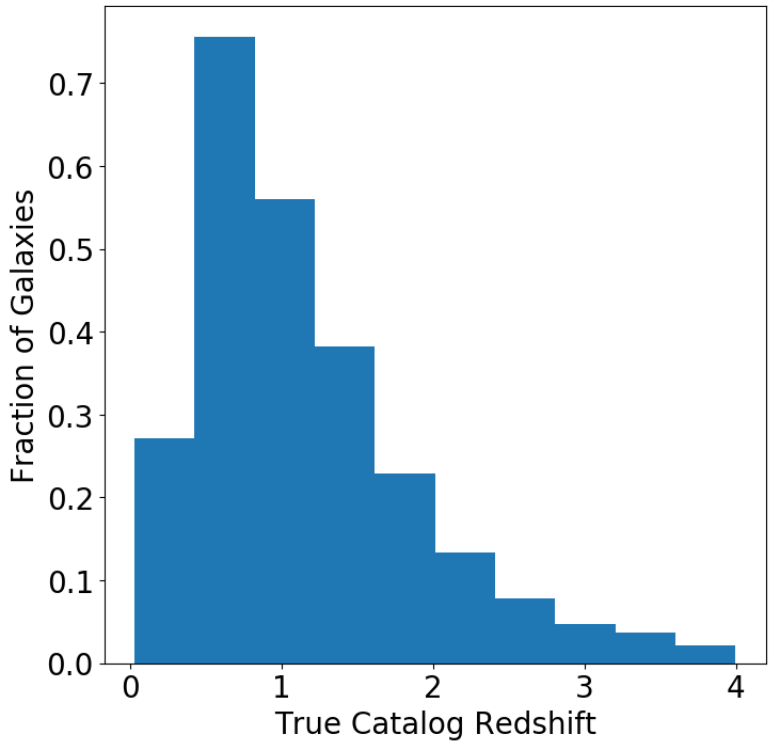
	pc1	pc2	pc3	pc4	pc5	pc6
u	0.7977	-0.5534	-0.2285	-0.0726	-0.004188	-0.001952
u-g	0.5969	0.7661	0.1783	0.1568	0.02073	0.0004168
g-r	0.08518	-0.1745	0.8555	-0.4118	-0.2461	-0.01691
r-i	0.01012	-0.2212	0.3991	0.4946	0.7049	0.2239
i-z	-0.006634	-0.1591	0.1573	0.6848	-0.4525	-0.5255
z-y	-0.003662	-0.04681	0.00884	0.2948	-0.4872	0.8206

1.1.1. *Features:*

Previous algorithms have had more success using a set of transformed features commonly called 'colors'. This transformation is motivated by physics, and is done by subtracting the measurements in adjacent bins, pairwise, bringing the 6 measurements down to 5 features. By including the original measurement from one bin (it doesn't matter which one, unless it carries an unusually large error), no information is lost in the transformation. So the final feature set includes 5 colors and 1 raw measurement for a total of 6 features.

We could also transform the measurement errors and include them as features. The simplest way to do this is to assume the errors are uncorrelated and add them in quadrature. Indeed, this is what is done in [3]. However, with one exception (see 2.1.3), I leave them out of the final dataset for two reasons. First, I'm not sure how the errors were calculated (this is simulated data), and it is possible that the information used to calculate the true redshift was also used to calculate these errors in a way that provides the algorithm with access to the true information, even in the test set. Second, there are multiple factors that could cause the errors to be correlated, so adding them in quadrature is not necessarily the best thing to do. A more thorough analysis of the errors in the dataset is an avenue for future work.

FIGURE 1. Histogram of true redshifts in the dataset, reproduced from [3]. True redshifts become more difficult to obtain as the redshift increases. Algorithms may have a more difficult time making accurate predictions at higher z because of the sparsity of training data.



1.1.2. *True Redshift (spec-z).* The calculation of the redshift from measurements of light generally depends on being able to find known features in the intensity as a function of frequency and measure how far they have been shifted along the spectrum. As a result, poor frequency resolution propagates to increased error in the redshift. This becomes important when we consider the two ways in which telescopes can take measurements: spectroscopy and photometry.

Spectroscopy records information about the amount of light coming in over a wide range of the frequency spectrum, at high resolution. Therefore, a redshift calculated from spectroscopic measurements is very precise and can be taken as the true redshift (sometimes called a 'spec-z').

Photometry essentially divides the spectrum into a small number of bins (usually on the order of 5) and records only aggregated information for each bin. Thus photometry is much cheaper to do and so we have (and LSST will be able to get) much more data of this type. This data can be collected quickly for large numbers of galaxies and used by ML algorithms to estimate the redshift (called a 'photo-z'). However, the low resolution necessarily leads to less accurate predictions. This provides further motivation to study how the errors scale with sample size for various algorithms.

2. METHODOLOGY

I follow [6] in their evaluation of photo-z estimates as a function of sample size. The traditional metric is the scaled difference:

$$\Delta z \equiv \frac{photo_z - spec_z}{1 + spec_z}. \tag{1}$$

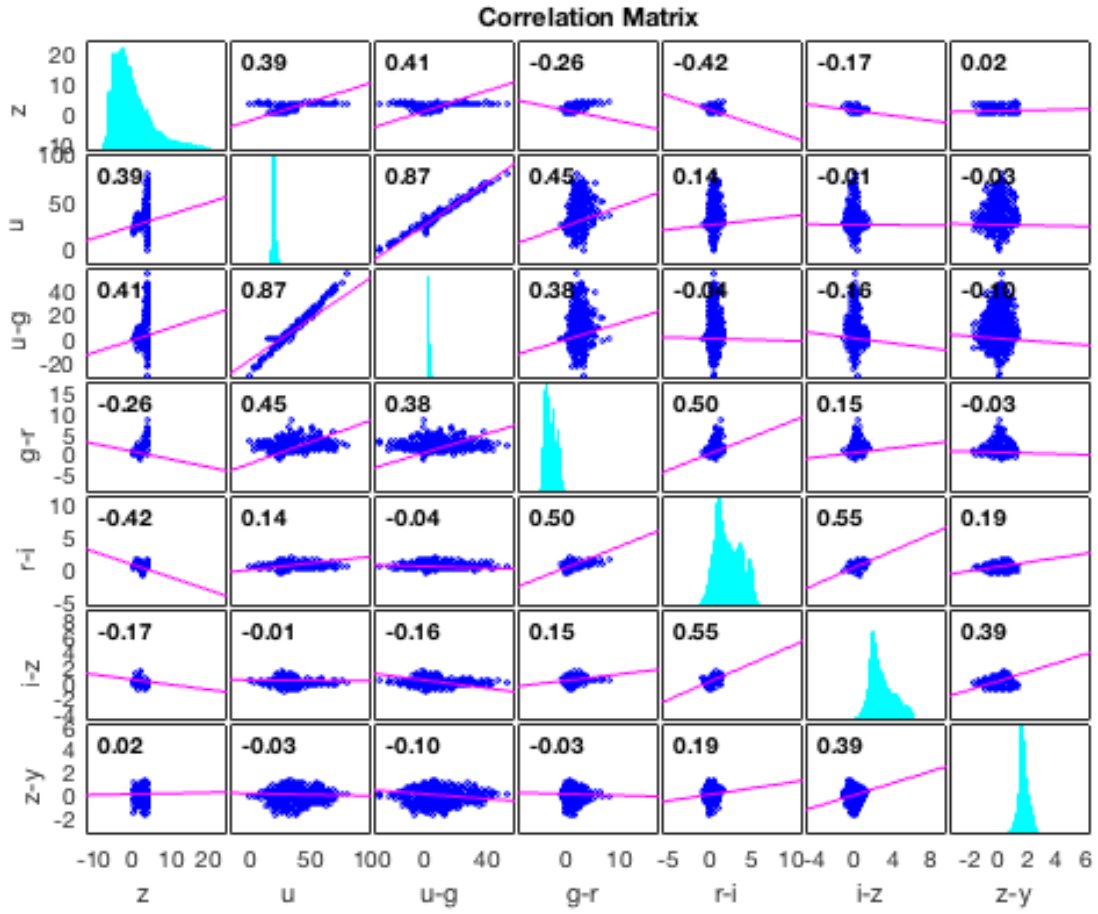
I evaluate the following two statistics on the metric:

$$\text{NMAD} = 1.48 \times \text{median}(|\Delta z|) \tag{2}$$

$$\text{OUT10} = \frac{1}{N} \sum_{n=1}^{N} \left[|\Delta z_n| > 0.1 \right] \tag{3}$$

NMAD is the normalized, median absolute deviation and OUT10 is the fraction of predictions for which $|\Delta z| > 10\%$. OUT10 is an important statistic because photo-z algorithms are prone to catastrophic errors in the

FIGURE 2. Pair-wise scatter plots with histograms along the diagonal. Correlation coefficients are printed on each plot.



predictions, due to both the physics involved and the inherently low resolution of photometry. To quantify the scaling, I fit these statistics to power laws of the form, $a + b \times N^c$, where N is the training sample size.

For the neural net (NN) and random forest (RF) main results I use a single test set of 10^5 (10^4 in smaller test runs, see algorithm subsections) randomly selected galaxies for the predictions, separated from the training sets prior to training. The GPz code handles the train/validate/test splits internally, but I set Ntest = 10^5 (10^4) to maintain some consistency.

Because I aim to evaluate the general performance of the algorithms, I train 20 ($\sim$ 4) models for each algorithm and training sample size and pool the results before calculating the statistics.

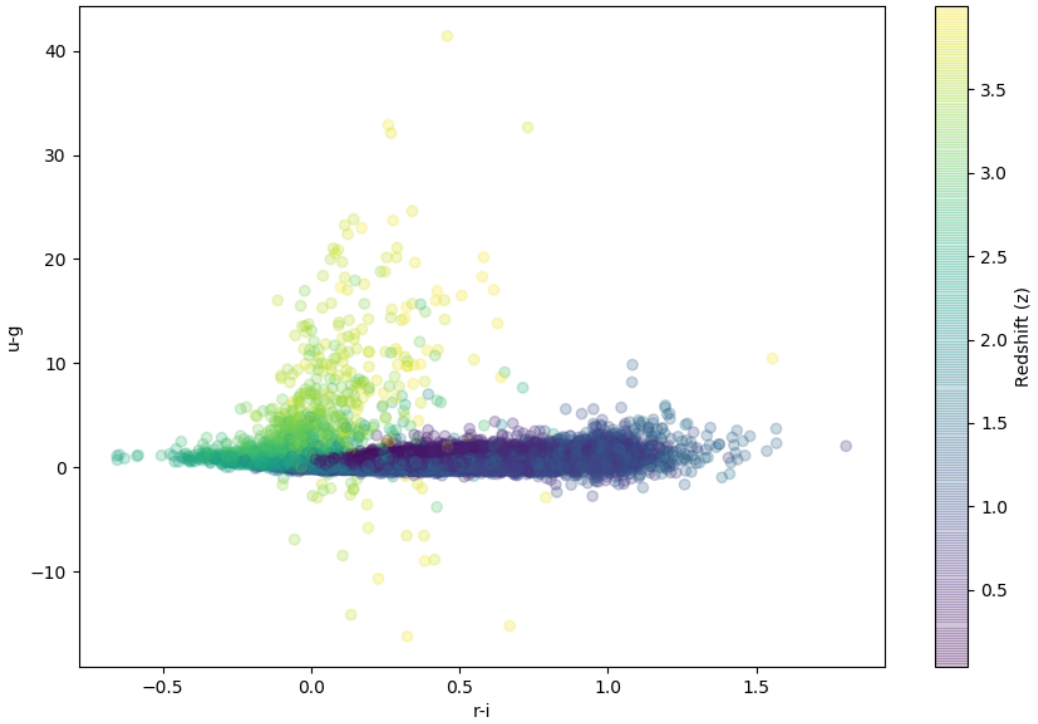
2.1. **Algorithms.** Various algorithms have been used on this problem with NNs and RF regressors being the most common. See, for example, [5], [7], and [4]. My main results (Figure 4) evaluate and compare the performance of four algorithms:

(1) NN composed of 2 hidden layers with 10 units each.
(2) NN composed of 3 hidden layers with 15 units each.
(3) RF regressor composed of bagged decision trees.
(4) GPz which is a publicly available code based on Gaussian Processes and kernel density estimation, developed specifically for photo-z's.

2.1.1. *Neural Nets.* I train multi-layer neural networks using two different architectures: `2x10` with 2 hidden layers, each with 10 units; and `3x15` with 3 hidden layers, each with 15 units. Both are motivated by approaches in [5] (see sections 4.1.1 DESDM and 4.1.2 ANNZ).

I use the Matlab `fitnet()` function with backpropagation optimized using the Levenberg-Marquardt method. After running a few tests I set the parameters `epochs=500`, `max_fail=50`, `min_grad=1e-10` and use the `tanh` transfer function.

FIGURE 3. Scatter plot of the two features with highest absolute value correlations with the redshift, colored by true redshift. Galaxies at low z tend to cluster near $u - g = 0$. At high z there is a much larger scatter in $u - g$ but the $r - i$ values tend to be lower. This is a random sample of 10000 galaxies to avoid saturating the plot. I verified that the plot looks qualitatively similar for different random samples.



2.1.2. *Random Forest Regression.* I train random forest regression models using the Matlab `fitrensemble()` function. I did some preliminary runs with `OptimizeHyperparameters='auto'` and found the following "best" options:

Method	Bag
NumLearningCycles	495
MinLeafSize	1

Guided by these results I tested some settings and ultimately use `Method='Bag'` with `Learners='tree'`, `MaxNumSplits=Nsamples-1`, and `NumLearningCycles=500`, `Crossval='off'`. This generates a random forest model using 500 weak learner decision trees, each trained on a subset of data of size N generated via bootstrap resampling.

2.1.3. *Gaussian Process Regression using GPz.* GPz is a publicly available code developed specifically for photo_z estimates. The method is described in [2], and the specific code is introduced in [1] and available at `https://github.com/OxfordML/GPz`. Since we studied this type of technique only briefly in the course I will outline the approach in a little more detail.

A Gaussian Process (GP) is a non-parametric, non-linear, regression algorithm. It assumes output, y, is predicted by some function of the input, $\mathbf{x}$, plus Gaussian noise $\epsilon \sim N(0, \sigma^2)$:

$$y = f(\mathbf{x}) + \epsilon.$$

Then the conditional probability of y given f is Normally distributed as $p(y|f) \sim N(f, \sigma^2)$ and Bayes' Theorem can be used to write

$$p(f|y, \mathbf{X}) = \frac{p(y|f)p(f|\mathbf{X})}{p(y|\mathbf{X})} \tag{4}$$

The prior, $p(f|\mathbf{X})$, is modeled non-parametrically (except for hyperparameters) using kernels that model the density around each input point. GPz uses radial basis functions for these kernels. Standard GP models are computationally expensive since there is a kernel for each datapoint and the solution requires us to invert an NxN covariance matrix associated with the kernels. GPz dramatically reduces the complexity by using sparse kernels and maintains performance by optimizing hyperparameters (governing the shape and length scale) that are unique to each kernel rather than standard, global hyperparameters. This allows kernels to specialize on different regions of parameter space, and this flexibility is cited as a key reason for the success of GPz (see [2] for a detailed derivation).

GPz also has a large component that is focused on estimating the variance in the prediction due to two factors, the input noise (errors on the measurements) and uncertainty due to the density of training points in a particular region. 'True' redshifts are more difficult to obtain at higher redshifts, and so the training data is not uniformly sampled (see Figure 1). I only use the point estimate of (4) in this work, so I will not describe the error calculations. However, this means that GPz requires the use of the measurement errors, so these models use more information than the NN or RF cases.

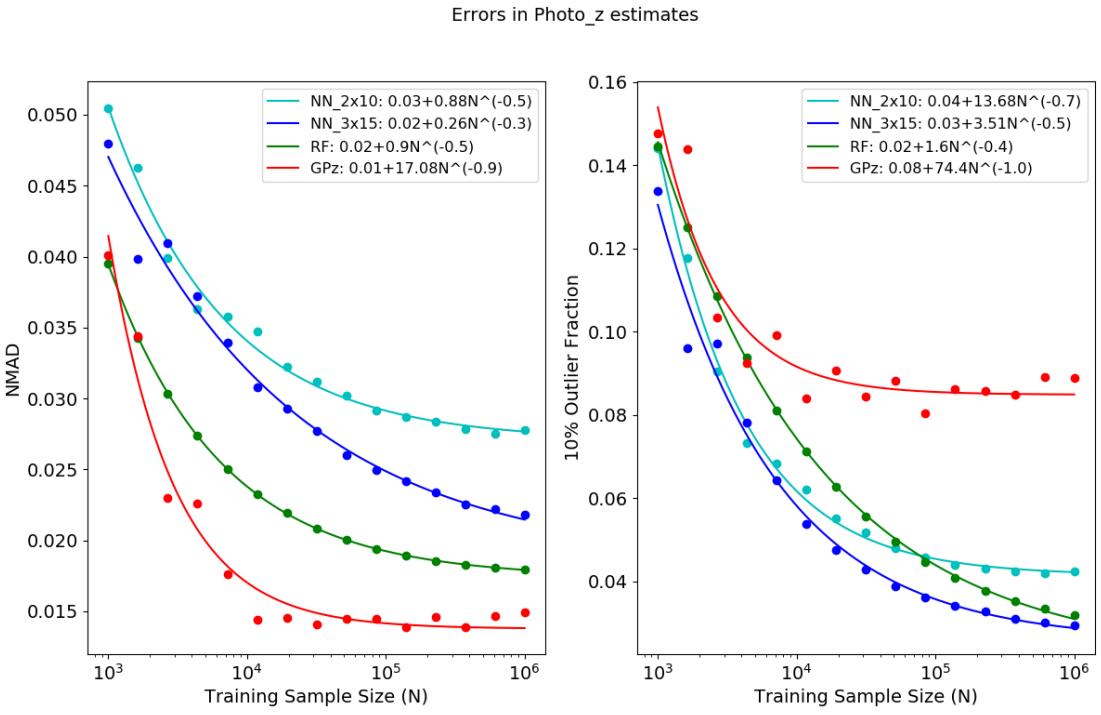
GPz performs the feature transformation internally, so the inputs to this algorithm are the raw measurements, including the errors. Additionally, it minimizes $|\Delta z|$ directly, so it can be expected to perform better than NN or RF on our problem.

I ran several small tests with different settings. I found two that improved the predictions and used them in my main results runs: increasing the maximum number of iterations, and using the errors as features (rather than including them in the prediction uncertainty). I chose to use the errors as features since the use of the errors is required one way or the other. I intended to do a more fair comparison with the NN and RF models by following the method used in [3] to add the error features to those datasets, but time constraints prevented this. In retrospect it would have been a better choice to use the errors in the variance estimates for the main runs.

3. Results

See Figure 4. The NN models tend to perform worse on the NMAD but better in OUT10. The RF results show the best convergence of any algorithm I tested, indicating that it may be the most stable for this problem. Comparing with [6] (who also tested RF), I find the same scaling of OUT10 ($N^{-0.4}$) and a slightly better scaling of NMAD ($N^{-0.5}$). GPz has the best scaling results ($N^{-0.9}$ and N^{-1}), but converges to the highest OUT10. The general behavior of limited improvement above $N \sim 10^4$ is also noted in [1]. Note also that GPz's use of the error information may be problematic and it has the most scatter around the power law fit.

Figure 4. Main results plotted as data points (pooled from 20 training instances). Power law fit shown as solid curves. See Section 3. Results from [6] are shown in Figure 5 for comparison.



In addition to the main results, I ran several smaller tests targeted at different settings:

FIGURE 5. Reproduced from Newman et al., 2019 ([6])

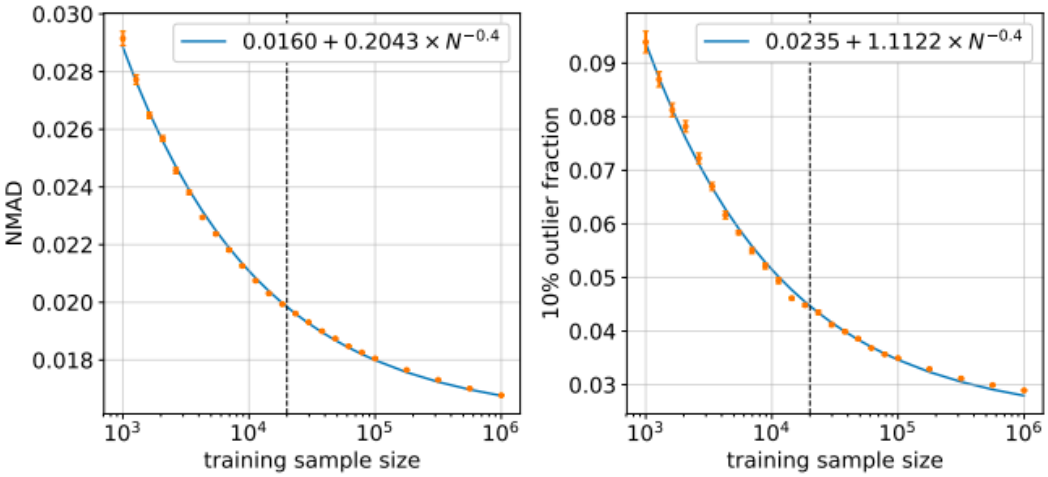


Figure 1: Orange points show photometric redshift errors and outlier rates versus the number of galaxies in the training set for galaxies with simulated LSST photometric errors. Photo-z's were calculated using a random forest regression algorithm. The left panel shows the photo-z error, quantified by the normalized median absolute deviation (NMAD) in $(z_{\rm phot} - z_{\rm spec})/(1+z_{\rm spec})$, as a function of training set size; similarly, the right panel shows the fraction of 10% outliers, i.e. objects with $|z_{\rm phot} - z_{\rm spec}|/(1+z_{\rm spec}) > 0.1$. A vertical dashed line shows the sample size for the baseline training survey from [5]. The blue curves represent simple fits to the measurements as a function of the training set size, N. This analysis uses a set of simulated galaxies from Ref. [11] that spans the redshift range of $0 < z < 4$, using a randomly-selected testing set of 10^5 galaxies for estimating errors and outlier rates.

3.1. **Neural Nets.** I did the following two NN tests, each with 6 sample sizes and 5 runs per sample size (see Figure 6):

(1) `3x50`: The predictions improve with the more complex network (`3x15`), so I ran a smaller test using 3 hidden layers with 50 units each to see if the trend continued. The results are a bit worse at small N and a bit better at large N, so this network improves more quickly as training sample size increases.
(2) `RELU` (`'poslin'` in Matlab) is a common transfer function in photo-z algorithms, so I tested this as well. This result is closer to the main results than `3x50`, but the fit is noisy, so it is hard to draw conclusions. More runs are needed here.

3.2. **GPz.** I tested the following three variants of GPz (see Figure 7):

(1) `Defaults` uses the measurement errors in the variance calculations rather than as additional input features.
(2) `input errors = 0` was trained on a modified dataset with all measurement errors set to zero.
(3) `'balanced' errors` weighs the rarer samples more heavily, attempting to account for the sparsity of data at high redshift. This was expected to be the weighting option that was best for this dataset, so its poor performance is surprising and worthy of further investigation.

Tests (1) and (2) attempt to put GPz on a more even footing with the NN and RF models.

GPz models are the most computationally expensive of those I tested, so I could not do as many runs. Therefore, this data is sparsely sampled in N and is pooled with only 2 runs per sample size, so no specific conclusions can be made. More training runs are needed.

4. CONCLUSIONS AND FUTURE WORK

I tested four machine learning algorithms on the problem of predicting the photometric redshifts of galaxies to understand how the errors in the predictions scale with training sample size. NN models tend to perform worse than the others in the NMAD, but result in fewer catastrophic errors (better OUT10). The RF model would probably continue to improve OUT10 with larger N ($a = 0.02$ in the fit), but acquiring more training samples

FIGURE 6. Results of neural net smaller test runs. `2x10` and `3x15` data is from main results. Note that `3x50` and `RELU` data is more sparsely sampled and is pooled from only 5 runs per model; separation of the data points from the fits suggests that results have not yet converged.

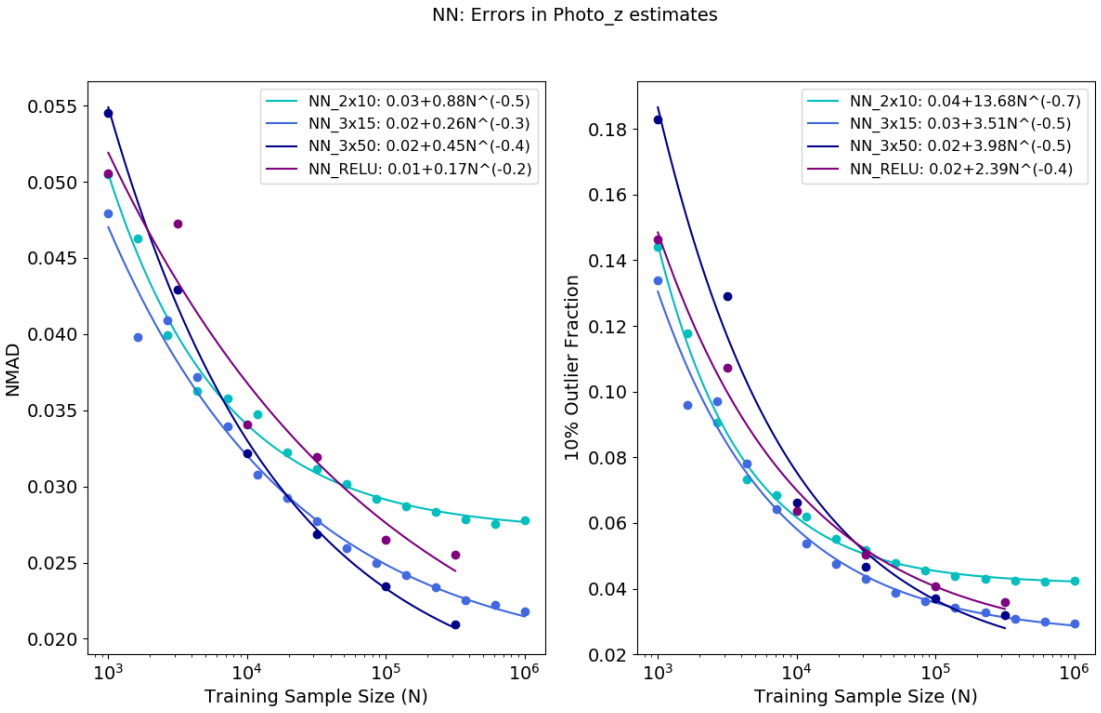
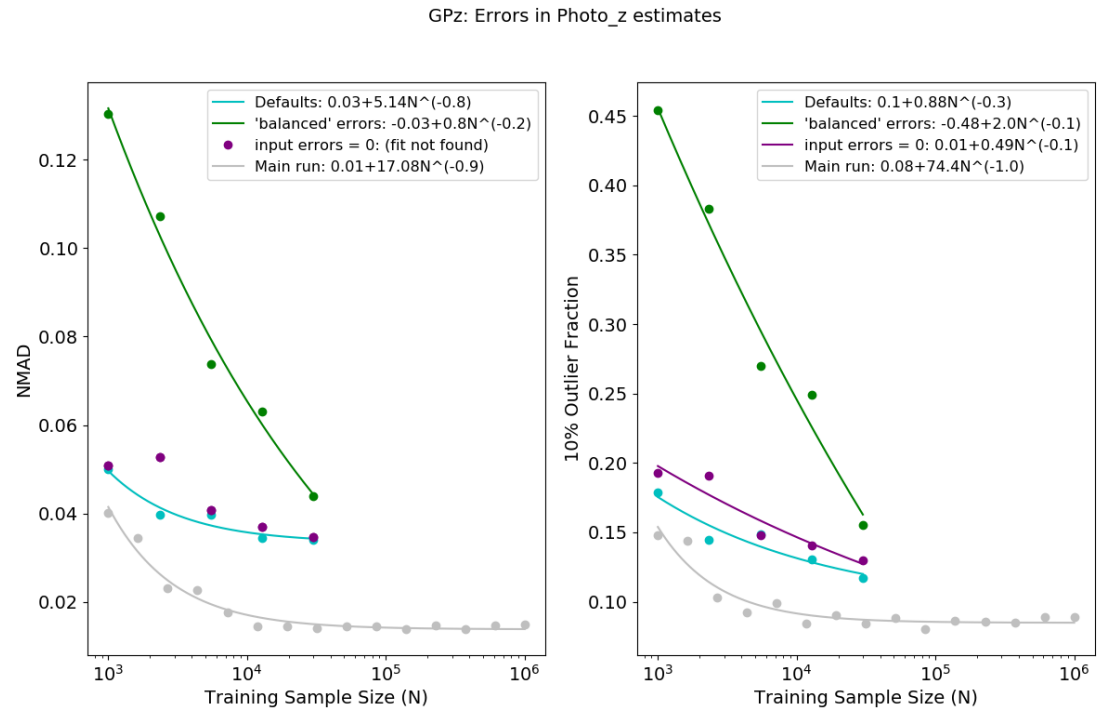


FIGURE 7. GPz smaller test runs. More runs are needed before drawing specific conclusions, but general trends can be seen. Refer to section 3.2 for details.



is unrealistic. GPz has the best NMAD results, but converges to the worst OUT10 result, just above 0.08, so it makes the most catastrophic errors.

Many things could be done to refine and extend this work. I list a few here:

- There are other codes written explicitly for the photo-z problem that could be tested and compared to the results in this work (e.g. ANNz2 [7], TPz [4]).

- For a more fair test of GPz and comparison with other models, the measurement error features should be treatment more carefully.
- Experiment with larger and smaller numbers of basis functions used by GPz. A larger number may result in better coverage of the parameter space and a lower OUT10. Looking at a small set of basis functions ($\sim 5 - 10$) corresponding to the kernels most important to the predictions could provide insight into galaxy types and distributions. These optimized kernel density estimates could be compared with two things:
  (1) Current physical models that predict specific types of galaxies and what their photometry measurements should look like as a function of redshift.
  (2) Results of unsupervised, clustering methods applied to the dataset which may provide insight on how many distinct types of galaxies there are as a function of redshift.
- Density estimation on the subset of the test data with $|\Delta z| > 10\%$ to see if there are localized regions of the parameter space that are not well predicted. These results could also be compared to (1) and (2) above to search for insight.

References

[1] I. A. Almosallam, M. J. Jarvis, and S. J. Roberts. GPZ: Non-stationary sparse Gaussian processes for heteroscedastic uncertainty estimation in photometric redshifts. *Monthly Notices of the Royal Astronomical Society*, 462(1):726–739, 2016.

[2] I. A. Almosallam, S. N. Lindsay, M. J. Jarvis, and S. J. Roberts. A sparse Gaussian process framework for photometric redshift estimation. *Monthly Notices of the Royal Astronomical Society*, 455(3):2387–2401, 2016.

[3] M. L. Graham, A. J. Connolly, Ž. Ivezić, S. J. Schmidt, R. L. Jones, M. Jurić, S. F. Daniel, and P. Yoachim. Photometric Redshifts with the LSST: Evaluating Survey Observing Strategies. 1, 2017.

[4] M. C. Kind and R. J. Brunner. TPZ: Photometric redshift PDFs and ancillary information by using prediction trees and random forests. *Monthly Notices of the Royal Astronomical Society*, 432(2):1483–1501, 2013.

[5] C. Lidman, H. Lin, D. Burke, C. Cunha, N. Greisel, J. Zuntz, A. Fausti, K. Glazebrook, G. Bernstein, J. Gschwend, D. Gerdes, M. A. G. Maia, C. Sánchez, S. S. Allam, A. Dey, E. Sánchez, H. T. Diehl, D. Finley, M. Lima, D. Capozzi, R. Scalzo, A. Sypniewski, S. Jouvel, E. Fernández, I. Sadeh, S. Seitz, J. L. Marshall, J. de Vicente, J. Frieman, A. Walker, P. Pellegrini, I. Sevilla-Noarbe, A. Roodman, M. Carrasco Kind, M. J. Childress, B. X. Santiago, R. C. Nichol, F. J. Castander, A. Carnero, R. Miquel, P. Doel, F. Ostrovski, T. Davis, A. Kim, A. Evrard, F. Valdés, S. Desai, J. Estrada, O. Lahav, C. Bonnett, G. Tarle, F. B. Abdalla, D. L. DePoy, E. Gaztanaga, J. P. Bernstein, R. L. C. Ogando, D. L. Tucker, M. E. C. Swanson, W. Hartley, A. Amara, F. Yuan, K. Honscheid, S. A. Uddin, N. Kuropatkin, T. Abbot, D. Thomas, B. Flaugher, M. M. Rau, R. Brunner, E. Buckley-Geer, L. A. N. da Costa, M. Sako, K. Kuehn, M. Banerji, D. Atlee, P. Martí, M. Makler, and R. C. Smith. Photometric redshift analysis in the Dark Energy Survey Science Verification data. *Monthly Notices of the Royal Astronomical Society*, 445(2):1482–1506, 2014.

[6] J. A. Newman, J. Blazek, N. E. Chisari, D. Clowe, I. Dell'Antonio, E. Gawiser, R. A. Hložek, A. G. Kim, A. von der Linden, M. Lochner, R. Mandelbaum, E. Medezinski, P. Melchior, F. J. Sánchez, S. J. Schmidt, S. Singh, and R. Zhou. Deep Multi-object Spectroscopy to Enhance Dark Energy Science from LSST. 2019.

[7] I. Sadeh. ANNz2 - Photometric redshift and probability density function estimation using machine-learning. *Proceedings of the International Astronomical Union*, 10:316–318, 2015.