DECEMBER 14, 2013
Preprint typeset using LaTeX style emulateapj v. 04/17/13

ACCURATE REDSHIFT ESTIMATION FROM PHOTOMETRIC COLORS

CHRISTOPHER DAVIS[1], DEVON POWELL[1], TONY LI[1]
[1]Kavli Institute for Particle Astrophysics and Cosmology; Physics Department, Stanford University, Stanford, CA 94305, USA
SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA;
cpd@stanford.edu, dmpowel1@stanford.edu, tonyyli@stanford.edu
(Dated: December 14, 2013)

1. INTRODUCTION

Accurate distance measurements are vital to observational cosmology. When we detect an object like a galaxy, a common proxy quantity for its line-of-sight distance (as well as "look-back" time) from us is its *cosmological redshift*: the shifting to redder wavelengths of the electromagnetic spectra of very distant objects (often galaxies) due to the expansion of the universe as light travels toward us. [1] Theoretical cosmological models make specific predictions about the large-scale structure and statistical properties of galaxies as the universe evolves, so precisely testing these models requires measuring the redshifts of a large population of galaxies.

To clarify the language used in this paper, redshift values are conventionally labeled z, and the higher an object's redshift is, the farther away it is from us, or alternatively, the further back in time we see it.[2] An object observed "at redshift zero" ($z = 0$) is seen in the nearby universe, or very close to the present day, while an object at $z \sim 1$ is seen in a universe that is approximately half its current age.

The most straightforward type of redshift estimate is *spectroscopic redshift*: one observes the object's detailed spectrum and measures the wavelength offset of recognized emission or absorption lines. Unfortunately, this is prohibitively expensive and time-consuming for large galaxy surveys, which will catalog billions of galaxies. These surveys instead perform photometry, which measures the object's total brightness in several broad bands of the electromagnetic spectrum, necessarily losing detailed spectral information (see Figure 1 for an illustration). However, by comparing photometric measurments to known spectra, one can determine a probability distribution for the redshift of individual objects. Redshifts thus estimated from broad-band color information alone are termed *photometric redshifts*, commonly referred to by the shorthand "photo-z." Machine learning (ML) methods are indispensable in obtaining such estimates with sufficiently low variance and bias.

An important note on units: brightnesses in astronomy are quoted in a reverse-logarithmic measure called *magnitudes*. Higher magnitudes m in a given photometric filter correspond to logarithmically *dimmer* brightnesses. Specifically, for two objects A and B with observed fluxes F_A and F_B

$$m_A - m_B = -2.5 \log_{10}(F_A/F_B) \quad (1)$$

so if A has a higher magnitude than B by 1, object A is in fact dimmer by a factor of $10^{2/5}$. While archaic, magnitudes are standard within the field and will be used in this study.

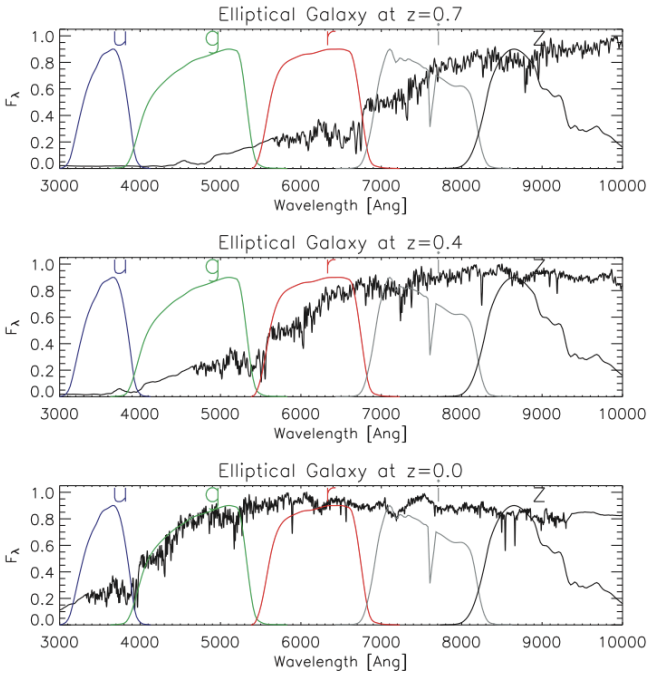

FIG. 1.— Illustration of the redshift of a galaxy spectrum. Each panel shows the simulated spectrum (*black*) of a galaxy observed at increasingly higher redshifts ($z = 0$, 0.4, 0.7). Also shown are the windows for five broadband photometric filters, which integrate out the detailed features present in the true spectrum. The goal of photometric redshift is to recover the redshift of a complicated galaxy spectrum from *only* the broadband data (i.e. one number per filter). From Padmanabhan et al. (2007).

2. DATA SET

We use a mock galaxy catalog ($\sim 1.4 \times 10^9$ objects) generated from the Aardvark simulation, one of several cosmological simulations run in preparation for the Dark Energy Survey by the SLAC National Accelerator Laboratory. This catalog includes simulated galaxies expected to be probed by the full depth of the Dark Energy Survey[3] (r-band magnitude of $\sim$23) and covers roughly a quarter of the sky (10,000 square degrees).

For each galaxy, magnitudes are provided in 5 broad wavelength bands: g (green, $\sim$ 550 nm), r (red, $\sim$ 650 nm), i, z, and Y (infrared, $\sim$ 800, $\sim$ 900, and $\sim$ 1000 nm respectively) bands. Together, these magnitudes comprise a 5-dimensional feature space. But because various systematic errors in photometric quantities can be ameliorated by considering the difference between magnitude bands (called *color*), we actually

[1] Redshift is a more directly observable quantity than distance or look-back time, because the exact conversion from redshift (which we observe) to distance (which we infer) depends on the cosmological model of the universe—the current rate of cosmological expansion, the universal proportions of dark matter and dark energy, etc. These parameters are estimated and well-constrained, but not known exactly.

[2] More precisely, if an object emits light at wavelength λ_i which is then observed at a longer wavelength λ_f, its redshift z is defined by $\lambda_f = \lambda_i(1+z)$.

[3] http://www.darkenergysurvey.org/

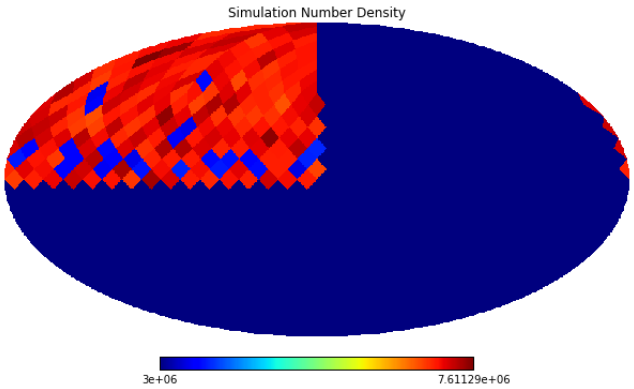

FIG. 2.— The distribution of objects on the sky, in Mollweide projection. Each Aardvark data file corresponds to one 54 square degree pixel on the sky.

use a 4-dimensional feature space comprising of the differences between wavelength bands, $(g-r, r-i, i-z, z-Y) \in \mathbb{R}^4$. From these, we ultimately wish to estimate a single value for each galaxy, the photometric redshift z_{photo}.

We use a small subset of galaxies selected to simulate a spectroscopic mini-survey of the DES footprint[4]. After applying reasonable magnitude and error cuts, we are left with $\sim 12,000$ galaxies. 70% are used to train and develop the various models, while the remaining fraction are saved to test the estimations.

3. METHODS

We apply the following machine learning algorithms to our data set with the end goal of performing a comparison between them. Further information on these techniques and others are reviewed in Zheng & Zhang (2012). Although various authors have applied these techniques to different data sets, we feel it is useful to have an 'apples-to-apples' look at the relative performance of these methods on the Aardvark catalog.

3.1. *SED Template Fitting (Non-ML technique)*

We first digress to comment on an extremely common non-ML technique for estimating photometric redshift. Spectral Energy Distribution (SED) templates estimate photometric redshifts by fitting some known model for an object type (i.e. quasar vs. elliptical galaxy vs. spiral galaxy) to the broad-band spectral magnitudes of an observed object. While SED templates are not strictly a machine learning technique, they merit a mention here because they can be used to divide data into subsets based on object type, giving Bayesian priors for ML techniques. Due to scope and time constraints, we will not use them in this project.

3.2. *Polynomial Regression*

Polynomial regression is a very basic technique which fits polynomial coefficients in $\mathbb{R}^c$ (the color space of c spectral magnitudes). Polynomial regression models of varying complexity (polynomial order, weighting scheme) have been applied since Connolly et al. (1995) with limited success.

[4] It is very important, however, to point out that any such spectroscopic survey will use a different instrument, possibly at a different site. The biggest consequence is that the population of galaxies from which the spectra are measured may in reality be significantly different from the photometric population. (This remains an issue even if one chooses for which galaxies to measure spectra.) We bypass this issue here.

3.3. *Spectral Connectivity Analysis*

This method uses a diffusion map to transform the training data from the raw color space into a space which encodes more relevant information through the diffusion distance (hence "connectivity analysis"). This is done by finding m principal eigenvectors of the $p \times p$ Gaussian kernel matrix (for p training samples) and forming transformed coordinates in $\mathbb{R}^m$. Because the method relies on the eigen-decomposition of a large matrix, it is necessary to estimate the eigenvectors using the Nyström method (cf. Press et al. 1992). Freeman et al. (2009) show that this method, combined with a weighted linear regression model, compares favorably to other ML techniques for photo-z estimation.

3.4. *k-Nearest Neighbors*

The k-nearest neighbors technique is an interpolation scheme between training data. The estimated redshift of a test data point is formed as a weighted sum of known redshifts from the k-nearest training examples. Here, "nearest" is intentionally left ambiguous, as it is possible to define different distance measures based on different (possibly nonlinear) coordinate transformations of the data space to find an optimal interpolation.

3.5. *Support Vector Regression*

The application of support vector machines (SVMs), a classification algorithm, to the linear space of photometric redshift estimates requires generalization to regression problems (see Smola & Schölkopf 2004) and is termed "support vector regression" (SVR). Analogous to the concept of support vectors in classification problems, the model produced by SVR only depends on a subset of the training data, because the cost function for building the model ignores any training data that is close to the model prediction. This method was first applied to redshift estimation by Wadadekar (2005), who noted that SVR is only useful when large training samples are available, which unfortunately is often not the case.

3.6. *Artificial Neural Networks*

Loosely categorized, various artificial neural network algorithms have also been successfully applied to estimating photometric redshifts. As with SVR, a number of studies (Collister & Lahav 2004; Firth et al. 2003; Abdalla et al. 2011) have noted that neural networks performed most competitively with other standard methods, such as SED Template Fitting, when a large representative training set was available, generally at mid-range redshifts, where a given section of the sky contains such a sample but is not too dim to be detected.

Zhang et al. (2009) and Vanzella et al. (2004) both included other object features, such as galaxy morphology, type classification, and size into their algorithms to yield increasingly accurate and computationally efficient estimates. Although naturally incorporated into their algorithms, the inclusion of additional such discriminating features will probably be beyond the scope of this project.

3.7. *Random Forest Decision Tree Regression (RFR)*

Each decision tree recursively partitions the training space such that samples with similar labels are grouped together. A each tree in the random forest is built from a bootstrapped sample of the training set. Each partition is chosen to be the best partition of a random subset of the features in training

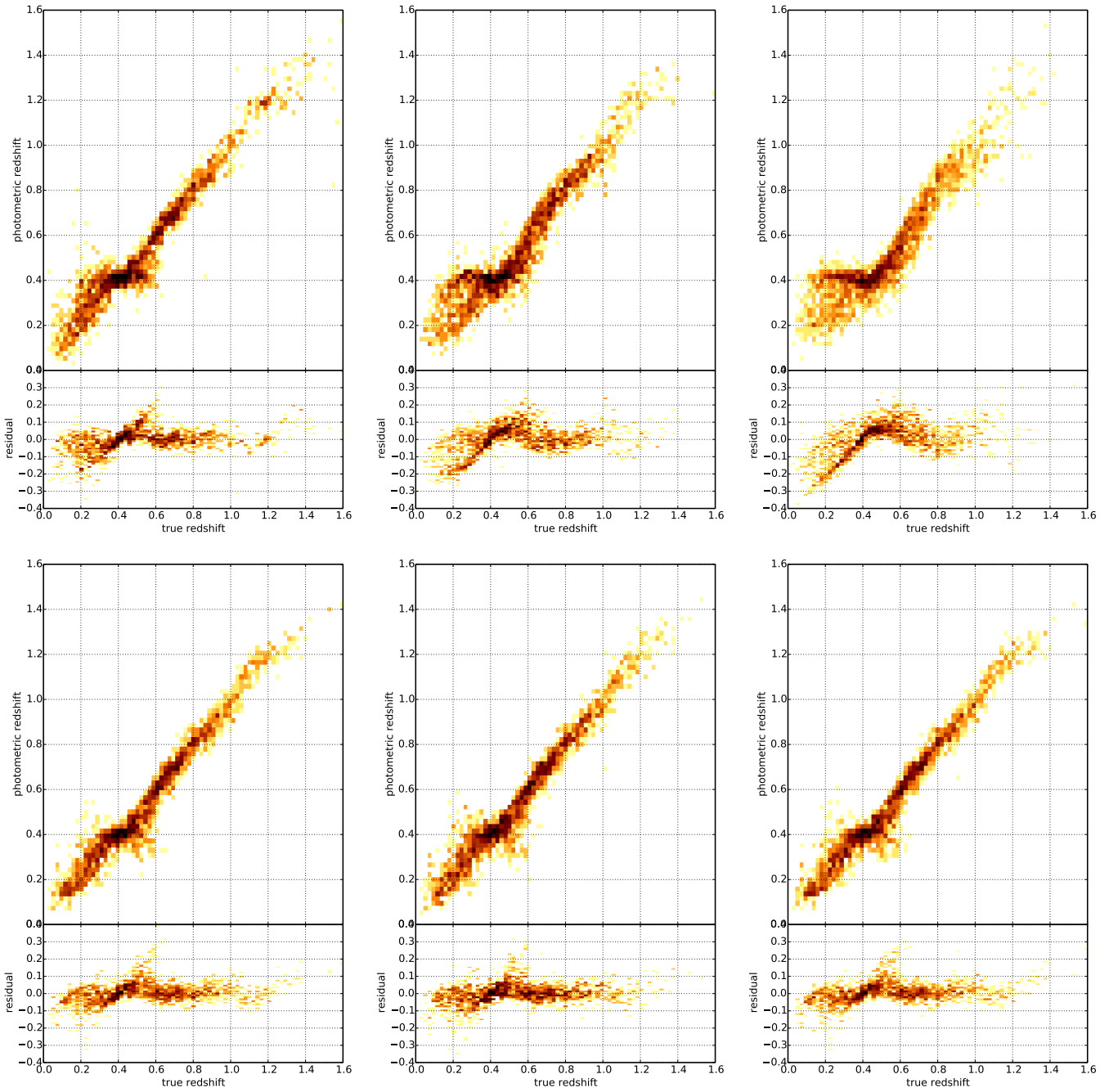

FIG. 3.— Estimated photometric redshift vs. true redshift, as obtained from the 6 methods presented in this study. Perfectly estimated redshifts would correspond to all points lying along a diagonal one-to-one line. Note the "kink" in the data around $z \sim 0.4$, which has a physical interpretation – see §5 for further discussion. **Top** (*left to right*): ANNz, PR, and SVR. **Bottom** (*left to right*): SCA, RFR, and kNN.

space. The estimated redshift is then the average of the decision trees. Similar methods have been implemented on real survey data using boosted decision trees (Gerdes et al. 2010).

4. RESULTS

Figure 3 displays our main results for each of the 6 methods we tested. For each method, we have plotted the photometric redshift estimate against the true redshift, as well as the residual (photometric redshift − true redshift).

Polynomial regression and support vector regression are the most naïve models we applied, yet they show quite good agreement between true and predicted redshifts.

The change of basis introduced in the spectral connectivity analysis appears not to have improved the errors on predicted redshift compared to vanilla *k*NN. This suggests that the photometric data lack sufficient latent connectivity to make a diffusion map useful for determining photometric redshifts.

k-Nearest Neighbors was nominally the best-performing method we applied. However, this may have been due to the abundance of training data (70/30 training/testing), giving a very well-filled parameter space with which to predict the redshift of test data

The artificial neural network was the worst performer. We tried several different neural net architectures, but failed to

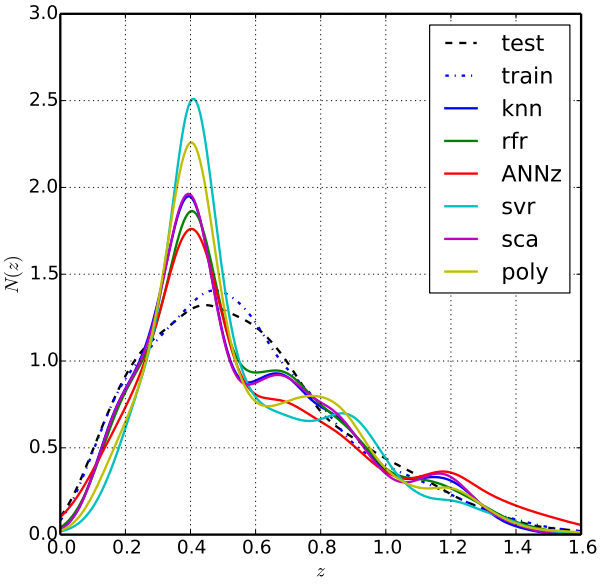

FIG. 4.— PDFs of photometrically estimated redshifts vs. the true distribution. Note the bias around $z = 0.4$ due to the 4000 Å break.

Method	σ_z
Artificial Neural Network (ANN)	0.2526
Polynomial Regression (PR)	
$d = 4$	0.1103
Support Vector Regression (SVR)	0.1229
Spectral Connectivity Analysis (SCA)	
$\epsilon = 0.1, m = 3$	0.0910
Random Forest Decision Tree Regression (RFR)	0.0966
k-Nearest Neighbors (kNN)	
$k = 30$	0.0893

TABLE 1
METHODS USED AND RMS ERRORS.

generate results competitive with our other methods. However, this method may still be viable given time to explore more complex architectures that can more robustly reproduce the regression model needed for redshift estimation.

At its heart, Random Forest Decision Tree Regression is similar to kNN (since it behaves like a *kd* tree in color space), so it is unsurprising that it performs comparably to kNN.

5. DISCUSSION, EXTENSIONS, AND CONCLUSIONS

There is a very apparent 'kink' in our results around redshifts of $z \sim 0.4$, which has a physical explanation. Here, photometric redshift predictions are relatively poor because a prominent spectral feature, the so-called "4000 Å break" (1Å = 10^{-10} m), quite literally "falls through the cracks" between the adjacent g and r filters, resulting in a degeneracy in color-redshift data. While other such degeneracies are present at higher redshifts, when the 4000 Åfeature falls between other filters, the design of the instrument is such that the gap between g and r filters is largest.

Because the growth of structure (manifested in the number distribution of galaxies as a function of redshift) is strongly affected by the underlying cosmology of our universe, it is useful to check the photometric redshift distribution against the spectral redshift distribution. We do so in 4, noting that the "4000 Å break" causes a significant bias at $z \sim 0.4$. While the neural network suffers the least from this bias, it instead significantly biases galaxies to higher redshifts.

The data set we applied these ML techniques to is quite optimistic in the context of real galaxy surveys. There is a selection bias towards bright galaxies at low redshift which makes obtaining sufficient training data for high-redshift objects (which are of most interest for constraining cosmological paramters) difficult. An additional benefit of our using simulated data is that we were able to use a 70/30 training/testing split, so that we have a large sample of data that is independent of the training. This is in large part responsible for the success of our photometric redshift estimates.

The main problem with photometric redshifts in the context of machine learning is that the training data exhibit very large noise. A huge boon for future redshift estimation techniques will be to improve the reparameterization of photometric information into optimal forms for machine learning algorithms. In addition, improvements to the fidelity of the training set (reducing incompleteness, noise, and misclassification) will greatly improve the accuracy and precision of photometric redshifts.

A possible future direction for this project could be to develop a model-generative machine learning approach which would produce a joint probability distribution in color magnitude – redshift space. This would be extremely useful, as it would allow for selective filtering of photo-z data in order to minimize uncertainties on constrraints of cosmological parameters.

We thank Prof. Risa Wechsler and Carlos Cunha for helpful discussions which guided the direction of this project, and we thank Michael Busha for providing access to the Aardvark DES mock galaxy catalogs.

REFERENCES

Abdalla, F. B., Banerji, M., Lahav, O., & Rashkov, V. 2011, MNRAS, 417, 1891
Collister, A. A., & Lahav, O. 2004, PASP, 116, 345
Connolly, A. J., Csabai, I., Szalay, A. S., et al. 1995, AJ, 110, 2655
Firth, A. E., Lahav, O., & Somerville, R. S. 2003, MNRAS, 339, 1195
Freeman, P. E., Newman, J. A., Lee, A. B., Richards, J. W., & Schafer, C. M. 2009, MNRAS, 398, 2012
Gerdes, D. W., Sypniewski, A. J., McKay, T. A., et al. 2010, ApJ, 715, 823
Padmanabhan, N., Schlegel, D. J., Seljak, U., et al. 2007, MNRAS, 378, 852
Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 1992, Numerical recipes in C. The art of scientific computing
Smola, A. J., & Schölkopf, B. 2004, Statistics and Computing, 14, 199
Vanzella, E., Cristiani, S., Fontana, A., et al. 2004, A&A, 423, 761
Wadadekar, Y. 2005, PASP, 117, 79
Zhang, Y., Li, L., & Zhao, Y. 2009, MNRAS, 392, 233
Zheng, H., & Zhang, Y. 2012, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 8451, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series