

THE EFFICACY OF GALAXY SHAPE PARAMETERS IN PHOTOMETRIC REDSHIFT ESTIMATION: A NEURAL NETWORK APPROACH

J. SINGAL¹, M. SHMAKOVA¹, B. GERKE¹, R.L. GRIFFITH², J. LOTZ³

Draft 2.3

ABSTRACT

We present a determination of the effects of including galaxy morphological parameters in photometric redshift estimation with an artificial neural network method. Neural networks, which recognize patterns in the information content of data in an unbiased way, can be a useful estimator of the additional information contained in extra parameters, such as those describing morphology, if the input data are treated on an equal footing. We show that certain principal components of the morphology information are correlated with galaxy type. However, we find that for the data used the inclusion of morphological information does not have a statistically significant benefit for photometric redshift estimation with the techniques employed here. The inclusion of these parameters may result in a trade-off between extra information and additional noise, with the additional noise becoming more dominant as more parameters are added.

Subject headings: techniques: photometric - galaxies: statistics - methods: miscellaneous

1. INTRODUCTION

Obtaining sufficiently accurate photometric redshift estimates is of the utmost importance for the current and coming era of large multi-band extragalactic surveys (see e.g. Huterer et al. (2006) for a recent review). Unlike spectroscopic redshift determination, photometric redshift estimation (photo-z) is highly subject to systematic errors and confusion because the spectral information of a galaxy is limited to the magnitude or flux in a number of wavelength bands.

Photo-z estimation techniques have traditionally been divided into two main classifications. So-called “Template fitting” methods, for example the popular *Lephare* package as described in Ilbert et al. (2006) and *Bayesian Photometric Redshift (BPZ)* as described in Benítez (2000), involve correlating the observed band photometry with model galaxy spectra and redshift, and possibly other model properties. In contrast, so-called “Empirical” or “Training set” methods, such as artificial neural networks (e.g. *ANNz*, Collister & Lahav 2004) and boosted decision trees (e.g. *BDT*, Gerdes et al. 2010), develop a mapping from input parameters to redshift with a training set of data in which the actual redshifts are known, then apply the mappings to data for which the redshifts are to be estimated. There are advantages and disadvantages to each class of methods. Template fitting methods require assumptions about intrinsic galaxy spectra or their redshift evolution, and empirical methods require the training set to be ‘complete’ in the sense that it is representative of the target evaluation population in bulk in all characteristics.

In regard to photo-zs, science goals such as using weak lensing for cosmology are most affected by the number of outliers - those objects whose estimated photo-zs are far from the actual redshifts. In general, data sets with bands extending into the mid-infrared (e.g. H, J, and K bands) have more accurate photo-z estimation and fewer outliers. However, most upcoming large surveys, such as the Large Synoptic Survey Telescope (LSST, Ivezić et al. 2008), will have optical and near-infrared data only.

It is a reasonable hypothesis that galaxy morphology and redshift are correlated in such a way that the addition of morphological information could improve photo-z estimation. Reasons include the larger frequency of mergers at higher redshifts, and, perhaps more importantly, the general evolutionary trend from spiral to elliptical shapes.

The inclusion of morphological parameters in photo-z estimation has been studied by Vince & Csabai (2006) and Way & Srivastava (2006). Both works use Sloan Digital Sky Survey (SDSS) data. Way & Srivastava (2006) consider several empirical methods and show marginal improvement for some methods with the addition of morphological information. Vince & Csabai (2006) claim an improvement of between 1 and 3 percent in the RMS error in photo-z determination, however it is not noted whether this result is significant and the method of photo-z estimation is not discussed. We note that SDSS galaxy photometric data is a bit unusual in the context of data that will be used to constrain cosmological parameters from surveys such as LSST, in that SDSS photometric data has a greater representation of nearby galaxies and thus fewer potential outliers.

In this work, we explore the efficacy of adding parameters describing the morphological information of galaxies, in the context of a neural network estimation technique for photo-zs.

2. DATASET AND SHAPE PARAMETERS

For this analysis we desire data with magnitudes in a number of optical bands, spectroscopic redshifts, and enough imaging resolution to determine morphologi-

Electronic address: jsingal@stanford.edu

¹ Kavli Institute for Particle Astrophysics and Cosmology
SLAC National Accelerator Laboratory and Stanford University
2575 Sand Hill Road, Menlo Park, CA 94025

² Jet Propulsion Laboratory
MS 169-327, 4800 Oak Grove Dr.
Pasadena, CA 91109

³ Leo Goldberg Fellow
National Optical Astronomy Observatory
950 North Cherry Avenue
Tucson, AZ, 85719

cal parameters. We use observations of the Extended Groth Strip from the the All-wavelength Extended Groth Strip International Survey (AEGIS) dataset (Davis et al. 2007), which contains photometric band magnitudes in u, g, r, i, and z bands from the Canada-France-Hawaii Telescope Legacy Survey (CFHTLS, Gwyn et al. 2008), imaging from the Advanced Camera for Surveys on the Hubble Space Telescope (HST/ACS, Koekemoer et al. 2007), and spectroscopic redshifts from the DEEP 2 survey using the DEIMOS spectrograph on the Keck telescope.

From the HST/ACS imaging data we form a set of parameters characterizing the morphological properties of the galaxies as follows:

1. The Concentration C: $C = 5 \log \frac{r_{80}}{r_{20}}$. This parameter defines the central density of the light distribution with radii r_{80} and r_{20} correspondingly 80% and 20% of the total light.

2. The Asymmetry A: $A = \frac{\sum_{x,y} |I_{(x,y)} - I_{180(x,y)}|}{2\sum_{x,y} I_{x,y}} - B_{180}$. This parameter characterizes the rotational symmetry of the galaxy’s light, with $I_{(x,y)}$ being the intensity at point (x,y) and $I_{180(x,y)}$ being the intensity at the point rotated 180 degrees about the center from (x,y). It is the difference between images rotated by 180°.

3. The Smoothness S: $S = \frac{\sum_{x,y} |I_{(x,y)} - I_{S(x,y)}|}{2\sum_{x,y} I_{x,y}} - B_S$. The smoothness is used to quantify the presence of small-scale structure in the galaxy. It is calculated by smoothing the image with a boxcar of a given width and then subtracting that from the original image. In this case $I_{(x,y)}$ is the intensity at point (x,y) and $I_{S(x,y)}$ is the smoothed intensity at (x,y). The residual is a measure of the clumpiness due to features such as compact star clusters. In practice, the smoothing scale length is chosen to be a fraction of the Petrosian radius.

4. The Gini coefficient G: $G = \frac{1}{\bar{X} n(n-1)} \sum_i^n (2i - n - 1) X_i$, describes the uniformity of the light distribution, with $G = 0$ corresponding to the uniform distribution and $G = 1$ to the case when all flux is concentrated in to one pixel. G is calculated by ordering all pixels by increasing flux X_i . \bar{X} is a mean flux and n is the total number of pixels.

5. M_{20} : $M_{20} = \log \Sigma M_i / M_{tot}$, is the ratio of the second order moment of the brightest 20% of the galaxy to the total second moment. This parameter is sensitive to the presence of bright off-center clumps.

6. The Ellipticity ε : $\varepsilon = 1 - \frac{b}{a}$. The values a and b are the semi-major axis and semi-minor axis of the galaxy.

The remaining two parameters are the fitting parameters to the Sersic profile form $\Sigma(r) = \Sigma_e e^{-k|(r/r_e)^{(1/n)} - 1|}$, where Σ_e is the surface brightness at radius r_e and k is defined such that half of the total flux is contained within r_e :

7. The Sersic power law index n and

8. r_e , the effective radius of the Sersic profile.

Morphological parameters C, A, S, G, and M_{20} are determined for these galaxies in Lotz et al. (2008), while ε , n , and r_e are determined by Griffith et al. (2011) using the Galfit package (Häusler et al. 2007). For this analysis we require a magnitude in each band, a spectroscopic redshift, and sufficient HST/ACS image reso-

TABLE 1
CONTENTS OF MORPHOLOGICAL PRINCIPAL COMPONENT VECTORS

...	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
C	-0.52	+0.13	+0.01	-0.17	-0.03	+0.19	0.09	-0.8
A	+0.10	-0.41	+0.62	-0.18	-0.61	+0.15	+0.07	-0.007
S	+0.06	+0.29	+0.72	-0.18	+0.60	-0.02	+0.04	+0.03
G	-0.47	-0.07	+0.09	-0.19	-0.09	-0.67	-0.51	+0.12
M_{20}	+0.49	+0.05	+0.04	+0.03	-0.09	-0.67	+0.32	-0.44
ε	+0.07	+0.62	-0.14	-0.63	-0.34	+0.03	+0.13	-0.23
n	-0.50	-0.006	+0.07	+0.21	-0.04	-0.20	+0.74	+0.32
r_e	-0.02	+0.58	+0.24	+0.65	-0.36	+0.01	-0.22	-0.03

lution to construct all 8 shape parameters. A total of 2612 galaxies spanning redshifts from 0.01 to 1.57 and i band magnitudes ranging from 24.43 to 17.62 are in the dataset used here.

Template-based photometric redshifts estimations for all of the galaxies used in this analysis have been reported in Ilbert et al. (2006). This estimation also provides a most likely galaxy type among template spectra corresponding to elliptical, Sbc, Scd, Irregular, or Starburst.

3. MORPHOLOGICAL PRINCIPAL COMPONENTS

In order to most efficiently determine the effect of the extra information provided by morphological parameters on the photo-z estimation, we form principal components of the morphological parameters. The morphological principal components are given as linear combinations of the eight morphological parameters discussed in §2 by Table 1.

Principal components (e.g. Jolliffe 2002) are the result of a coordinate rotation in a multi-dimensional space of possibly correlated data parameters into vectors with maximum orthogonal significance. The first principal component is along the direction of maximum variation in the data space, the second is along the direction of remaining maximum variation orthogonal to the first, the third is along the direction of remaining maximum variation orthogonal to both of the first two, and so on.

Given the available galaxy type estimations discussed at the end of §2, we can check for correlations between principal components of the morphological parameters and galaxy type, as in Figure 1. It is seen that the first principal component is well correlated with galaxy type, and correlations persist through several of the other principal components. These correlations indicate that the morphology may provide an additional handle on the photo-z estimation, since outliers often occur because a spectral feature (such as a break) of one galaxy type at a given redshift may be seen by the observer to be at the same wavelengths as a spectral feature of another galaxy type at different redshift. Thus morphological information indicative of galaxy type may help break this degeneracy.

4. PHOTO-Z ESTIMATION METHOD

Artificial neural network techniques have been popular empirical methods for photo-z estimation, including with such software packages as ANNz (Collister & Lahav 2004). In the case of neural network photo-z determination, the network functions as a ‘black box’ which finds patterns contained in the relation between band magnitudes (and, in principle, other information) and redshift in an unbiased way. Thus, a neural network photo-z

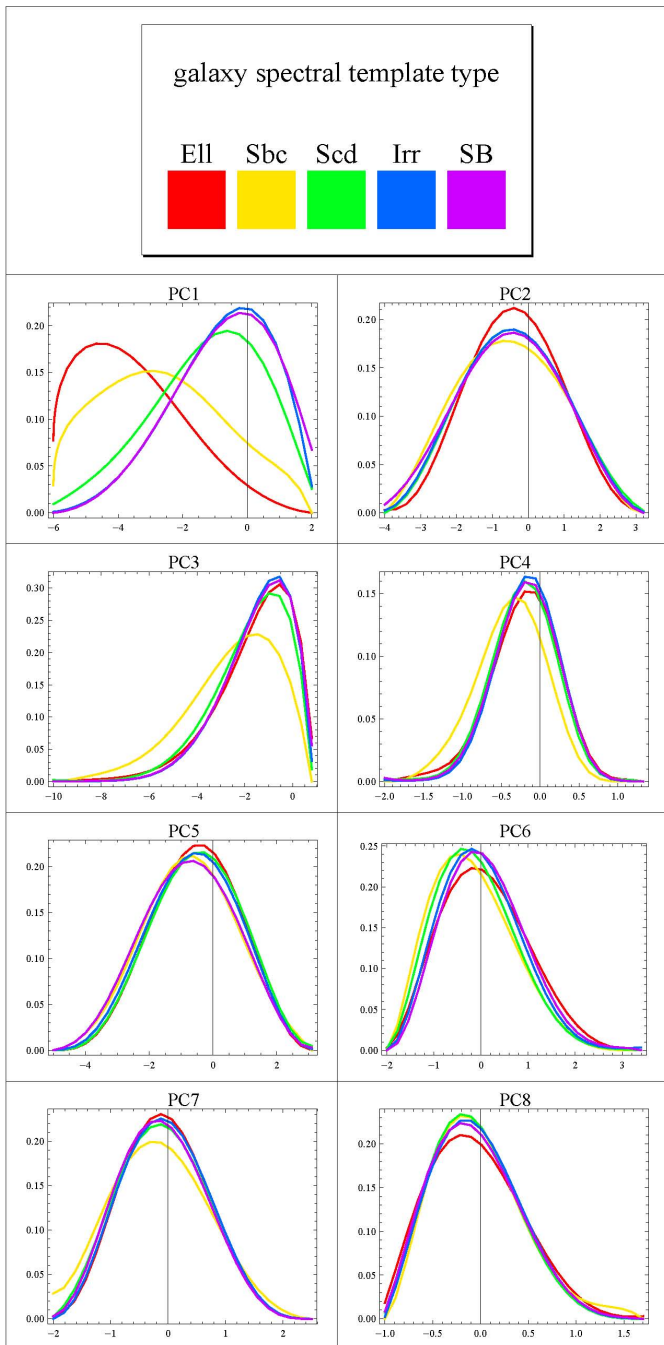


FIG. 1.— Distribution of morphological principal component values for galaxies of different type, for the 2612 galaxies used in this analysis. Galaxy type is estimated by Ilbert et al. (2006).

estimation can be a useful tool to explore whether additional parameters beyond the band magnitudes, such as morphology in this case, provide additional useful information.

An artificial neural network, in analogy with a biological one, contains layers of nodes called “neurons” and relationships between neurons in different layers of varying weights which can be altered. Each neuron in the network beyond the input layer assumes a value determined by passing through an activation function the sum of the product of all of the values of the neurons feeding the neuron in question times the weight of the connection between the two neurons. Neurons at the input accept

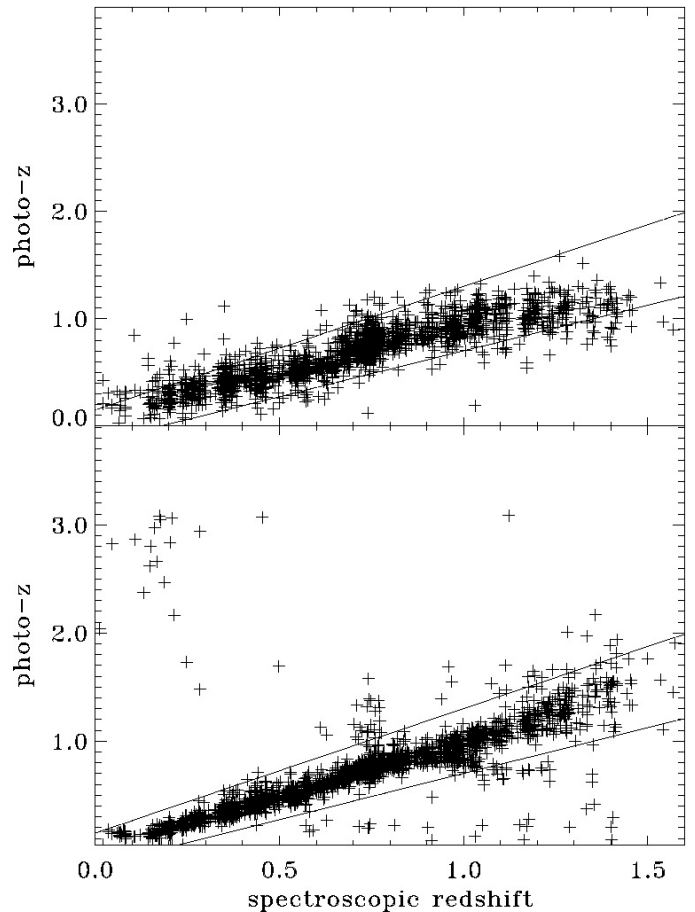


FIG. 2.— **TOP:** The estimated photo-z versus the actual redshift, as determined by the custom artificial neural network used in this work, for the case of no morphological parameters included. This determination is of the type used in this analysis, with 350,000 training iterations, where the training set is formed from 700 galaxies and the evaluation set, for which the results are plotted, consists of the remaining 1912 galaxies. ‘Outliers’ in a determination are defined as those where $|z_{phot} - z_{spec}| / (1 + z_{spec}) > .15$, shown as the two diagonal lines.

BOTTOM: The photo-z for the same galaxies as in the top plot, as estimated with the Lephare template fitting code as reported in Ilbert et al. (2006). The template fitting method has a lower scatter for non-outliers but a larger number of catastrophic outliers than the custom neural network for these galaxies.

values of data and the output of the network is the value of one or more output layer neurons.

The weights between neurons are adjusted by ‘training’ the network to best give desired outputs for a set of inputs. Training is dependent on a *training set* containing a number of cases with inputs and output(s). In the case of photo-z estimation, the training set contains band magnitudes and possibly other information as inputs, and the actual known (spectroscopic) redshifts as the output. With the weights set in this way, the network can be used to estimate the redshift of other galaxies in an *evaluation set*, and the results can be compared to the known redshifts of the evaluation set to determine the quality of photo-z estimation. A comprehensive discussion of artificial neural networks is presented in Haykin (1999).

The artificial neural network package used in this analysis is a ‘multi-layer perceptron’ developed for the IDL

environment by one of the authors (JS).⁴ Perceptrons are standard artificial neural network architectures for pattern recognition, consisting of input, hidden, and output neurons as described above. The primary motivation for the development of this code was to treat additional available galaxy information beyond photometric data (for example shape parameters) on an equal footing with the photometric data. The IDL code can be relatively easily modified, and could in principle be configured for a wide variety of input data situations. As training convergence is relatively slow in this network, it is most useful in situations where a robust training set is available from the outset.

As implemented here, the network has an input layer of neurons, five of which accept the observed magnitudes in each optical band, and an additional variable number of input neurons which accept values of as many morphological parameters as desired. The input layer treats all input information on an equal footing, normalizing each input parameter across all objects in the training set so that the inputs for each neuron on the input layer are distributed between 0 and 1. There are two hidden layers of 30 neurons each, and an output layer with a single neuron obtaining a value between 0 and 1 which is a proxy for the estimated redshift, with the linear conversion defined during the training when the known redshifts of the training set are supplied subject to the conversion.

The network uses a hyperbolic tangent activation function for the neurons beyond the input layer, and the weights are adjusted during training via the back propagation technique (e.g. Haykin 1999) where in each training iteration the weights are altered in a way to move ‘downhill’ in the high dimensional surface of summed training set redshift errors in the space of weights. Each iteration during training consists of the network evaluating the entire training set and adjusting the weights. In addition to standard back propagation, this network features an algorithm to ‘kick’ the weights away from possible local minima in the summed error. The top panel of Figure 2 shows the estimated photo-z versus spectroscopic redshift for the galaxies in the evaluation set of a particular determination with no morphological information included. This determination features 350,000 training iterations, 700 galaxies in the training set, and 1912 galaxies in the evaluation set, which is the standard used in all determinations here. The bottom panel of Figure 2 also shows photo-z estimations for the same galaxies with the Lephare template fitting method as reported in Ilbert et al. (2006). The custom neural network determination apparently leads to fewer catastrophic outliers than with the Lephare template fitting method, although has a larger scatter for those galaxies in which the photo-z estimate is close to the actual redshift.

5. RESULTS ON INCLUSION OF SHAPE PARAMETER INFORMATION

To determine the effect of including a given or multiple principal components in addition to the band magnitudes, we complete six realizations of the training and evaluation process for every case, with a training set of 700 galaxies with 350,000 training iterations and an eval-

uation set of the remaining 1912 galaxies, and record the number of outliers, and the RMS error, in the evaluation set. In this work we follow convention (e.g. Ilbert et al. 2006) and define outliers in a given realization as those galaxies where

$$\text{Outliers} : \frac{|z_{\text{phot}} - z_{\text{spec}}|}{1 + z_{\text{spec}}} > .15,$$

where z_{phot} and z_{spec} are the estimated photo-z and actual (spectroscopically determined) redshift of the object respectively. The RMS photo-z error in a realization is given by a standard definition

$$\sigma_{\Delta z/(1+z)} \equiv \sqrt{\frac{1}{n_{\text{gals}}} \sum_{\text{gals}} \left(\frac{z_{\text{phot}} - z_{\text{spec}}}{1 + z_{\text{spec}}} \right)^2},$$

where n_{gals} is the number of galaxies in the evaluation set and \sum_{gals} represents a sum over those galaxies. Note that we do not exclude outliers from the calculation of the RMS photo-z error. Because in each realization the membership of the training set varies, and because as the training process contains ‘kicks’ to knock the weights away from local minima in the summed error (see §4), each realization for a given input parameter set produces a slightly different number of galaxies in the evaluation set with outlier level errors, and a slightly different average error. For comparison, the template fitting results reported by Ilbert et al. (2006) give 5% outliers and an RMS error of $\sigma_{\Delta z/(1+z)} = .1881$ for this sample. This error is dominated by the catastrophic outliers (Figure 2), and drops substantially to below that of the custom neural network method if outliers are excluded.

Figure 3 shows the number of outliers and RMS error for the inclusion of the seven different principal components individually. Figure 4 shows the number of outliers and RMS error for the inclusion of multiple principal components, starting with none, then adding in the first, then adding in the first and second, then adding in the first through third, and so on. In each figure, the error bars correspond to the standard deviation of the number of outliers or RMS scatter in the different realizations. We note that with six realizations per case, the standard deviations in the number of outliers and RMS photo-z error are not particularly robust, however we include them to provide a sense of the scatter of results from different realizations. We note that the last principal component (PC8) should by definition contain minimal significant variation in the morphological parameters, so we do not include it in the analysis.

It is apparent that adding in any one of the principal components of the morphological parameters may provide a small decrease in the average number of outliers or the RMS error or both, but the differences are not statistically significant compared to the inclusion of no morphological information. Principal component 3 (PC3) seems to provide the most decrease in both outliers and RMS error, but still not enough to be considered statistically significant. As seen in Figure 4, adding multiple principal components increases the number of outliers and the RMS error.

6. DISCUSSION

⁴ available from www.slac.stanford.edu/~jacks

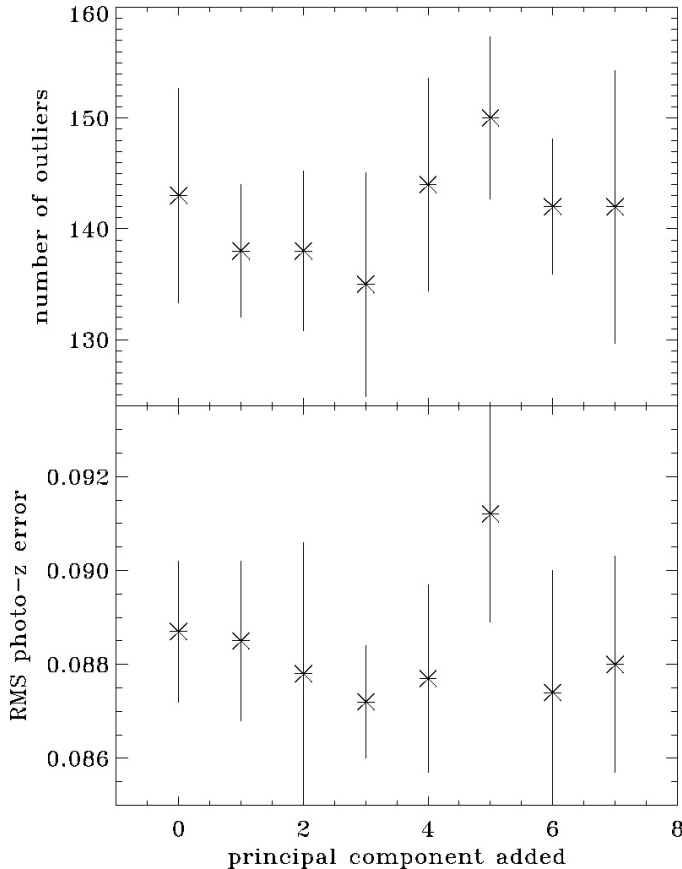


FIG. 3.— Number of outliers (TOP) and RMS error $\sigma_{\Delta z/(1+z)}$ (BOTTOM) in the photo-z estimation with the inclusion of different individual principal components of the morphological parameters. The uncertainties represent the standard deviation of the values obtained from different realizations, as discussed in §5.

We have used a custom artificial neural network for photometric redshift estimation to evaluate the effects of including galaxy shape information, in the form of principal components of morphological parameters, on the photo-z estimation. The dataset we use consists of 2612 galaxies with 5 optical band magnitudes, reliable spectroscopic redshifts, and eight morphological parameters each. We note that a neural network is in a sense an unbiased way of determining the relative strength of the correlations of a set of input parameters with the output parameter, and that this network is designed to treat all input parameters on an equal footing.

In order to more effectively include the morphological information, we form principal components of the morphological parameters. An analysis of the principal components and galaxy types shows that the value of the first few principal components and galaxy type are correlated.

However, we find that the inclusion of morphological information does not significantly decrease the number of outliers or RMS error in photo-z estimation for this dataset with the neural network technique used. When only one principal component of the morphological parameters is included, there can be a slight but not significant decrease. When multiple principal components of the morphological parameters are included, the number of outliers and the RMS error increases. We conclude that any gain that may arise in a neural network photo-z determination from correlations between morphology

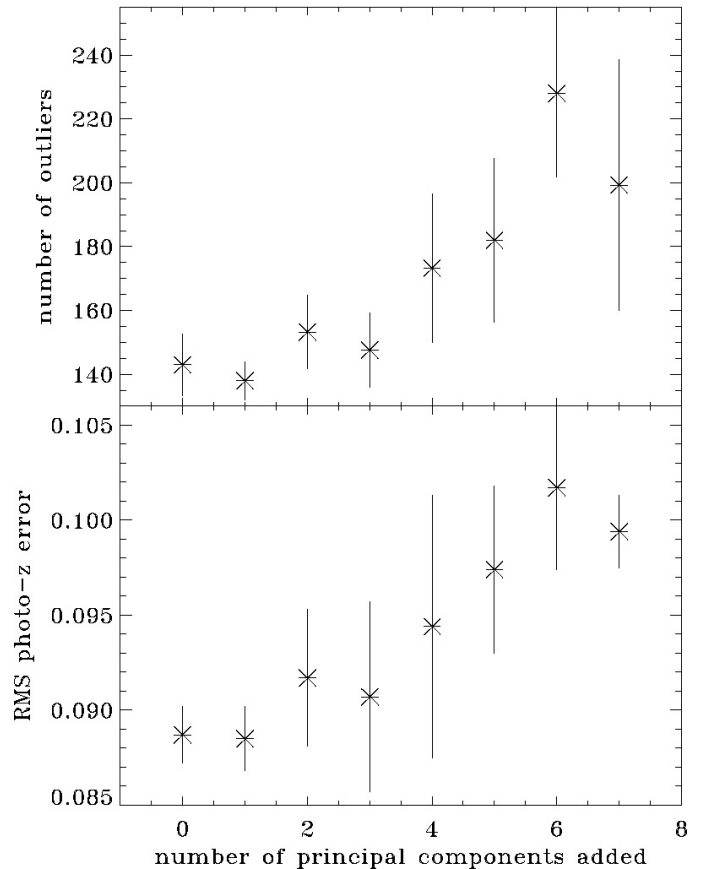


FIG. 4.— Number of outliers (TOP) and RMS error $\sigma_{\Delta z/(1+z)}$ (BOTTOM) in the photo-z estimation with the inclusion of multiple principal components of the morphological parameters, starting with none, adding the first morphological principal component, then the first and the second principal components, and so on. The uncertainties represent the standard deviation of the values obtained from different realizations, as discussed in §5.

and redshift in this dataset is overwhelmed by the additional noise introduced. It may be that any correlations between principal components of the morphological parameters and the galaxy type are degenerate to some extent with the correlations between galaxy type and galaxy colors.

This analysis is applicable to artificial neural network photo-z estimations, and possibly other training set methods, with similar data. It is possible, however, that morphological parameters could yield improvements for other algorithms, especially template-fitting methods with relatively large outlier fractions. This is because such outliers usually occur when a particular spectral break is confused for another break in a different galaxy type at a different redshift (e.g. an elliptical galaxy at low redshift mistaken for a spiral at high redshift). Having additional information to guide the template selection might therefore be helpful in reducing the outlier fractions. An analysis of the effects of shape parameters in photometric redshift estimation with a template fitting method will be presented in a forthcoming work.

JS thanks T. Brookings for his counsel, and S. Kahn and R. Schindler for their encouragement and support. MS is is thankful to R. Blandford and P. Marshall for very useful discussions and support. This work was supported in part by the U.S. Department of Energy under

contract number DE-AC02-76SF00515.

REFERENCES

- Benítez, N. 2000, ApJ, 536, 571
Collister, A. & Lahav, O. 2004, PASP, 116, 345
Davis, M., et al. 2007, ApJ, 660, L1
Gerdes, D., et al. 2010, ApJ, 715, 823
Griffith, R., et al. 2011, in prep
Gwyn, S., 2008, PASP, 120, 212
Häusler, B., et al. 2007, ApJS, 172, 615
Haykin, S., *Neural Networks: A Comprehensive Foundation*,
Upper Saddle River, NJ: Prentice Hall 1999
Huterer, D., Takada, M., Bernstein, G., & Jain, B. 2006, MNRAS,
366, 101
Ilbert, O., et al. 2006, A&A, 457, 841
Ivezic, Z. et al. 2008 arXiv:0805.2366
Jolliffe, I., *Principal Component Analysis*, Series: Springer Series
in Statistics, 2nd ed. New York, NY: Springer 2002
Koekemoer, A., et al. 2007, ApJS, 172, 196
Lotz, J., et al. 2008, ApJ, 672, 177
Vince, O. & Csabai, I. 2006, 'Toward more precise photometric
redshift estimation.' Proceedings of the International
Astronomical Union, 2, pp 573-574
Way, M. & Srivastava, A. 2006, ApJ, 647, 102