Accepted to ApJ
Preprint typeset using LaTeX style emulateapj v. 10/09/06

arXiv:0804.3413v1 [astro-ph] 21 Apr 2008

ROBUST MACHINE LEARNING APPLIED TO ASTRONOMICAL DATASETS III: PROBABILISTIC PHOTOMETRIC REDSHIFTS FOR GALAXIES AND QUASARS IN THE SDSS AND GALEX

Nicholas M. Ball[1], Robert J. Brunner[1,2], Adam D. Myers[1],
Natalie E. Strand[3], Stacey L. Alberts[1], David Tcheng[2]

Accepted to ApJ

ABSTRACT

We apply machine learning in the form of a nearest neighbor instance-based algorithm (NN) to generate full photometric redshift probability density functions (PDFs) for objects in the Fifth Data Release of the Sloan Digital Sky Survey (SDSS DR5). We use a conceptually simple but novel application of NN to generate the PDFs—perturbing the object colors by their measurement error—and using the resulting instances of nearest neighbor distributions to generate numerous individual redshifts. When the redshifts are compared to existing SDSS spectroscopic data, we find that the mean value of each PDF has a dispersion between the photometric and spectroscopic redshift consistent with other machine learning techniques, being $\sigma = 0.0207 \pm 0.0001$ for main sample galaxies to $r < 17.77$ mag, $\sigma = 0.0243 \pm 0.0002$ for luminous red galaxies to $r \lesssim 19.2$ mag, and $\sigma = 0.343 \pm 0.005$ for quasars to $i < 20.3$ mag. The PDFs allow the selection of subsets with improved statistics. For quasars, the improvement is dramatic: for those with a single peak in their probability distribution, the dispersion is reduced from 0.343 to $\sigma = 0.117 \pm 0.010$, and the photometric redshift is within 0.3 of the spectroscopic redshift for $99.3 \pm 0.1\%$ of the objects. Thus, for this optical quasar sample, we can virtually eliminate 'catastrophic' photometric redshift estimates. In addition to the SDSS sample, we incorporate ultraviolet photometry from the Third Data Release of the Galaxy Evolution Explorer All-Sky Imaging Survey (GALEX AIS GR3) to create PDFs for objects seen in both surveys. For quasars, the increased coverage of the observed frame UV of the SED results in significant improvement over the full SDSS sample, with $\sigma = 0.234 \pm 0.010$. We demonstrate that this improvement is genuine and not an artifact of the SDSS-GALEX matching process.

Subject headings: methods: data analysis — catalogs — quasars: general — cosmology: miscellaneous

1. INTRODUCTION

Advances in CCD and other technologies are enabling modern wide-field surveys to provide high quality photometry for ever-increasing numbers of astronomical objects (e.g., Kron 1995; Reshetnikov 2005; Lawrence 2007). Comparable advances in multifiber spectrographs are enabling similarly increasing numbers of spectra to be taken (e.g., Lahav & Suto 2004; Yip 2007). However, due to the increased integration time required to obtain a meaningful spectrum to a given depth compared to an image, the number of spectra available typically lags the number of images by more than an order of magnitude. Given the importance of the physical information contained within a spectrum compared to that within an image, for example, much more accurate diagnostics of an object's type and its redshift, any comparable information that can be obtained from the image is of great importance. With the much larger numbers of objects for which photometry is available, for applications that do not require high resolution spectra this information can even surpass the spectra in terms of statistical significance (e.g., Blake & Bridle 2005).

However, for many applications, it is vitally important to know not only the photometric information, but also its relative accuracy within the dataset for each object. Typically, this might be achieved by, for example, providing an error on a measured magnitude, or an estimated Gaussian dispersion on a photometric redshift. In general, however, one would like to utilize the full *probability density function* (PDF) within analyses, so that one can exclude objects which do not meet specific criteria, or fold the information into the analysis.

An area in which PDFs are of particular utility is photometric redshifts. For many purposes, provided that they are reasonably accurate, the final raw accuracy of a redshift estimate is not vitally important, provided that the error distribution is well known. Photometric redshifts, particularly those of quasars, are known to suffer from a percentage of 'catastrophic' failures, e.g., Budavári et al. (2001); Richards et al. (2001); Weinstein et al. (2004); Wu et al. (2004), in which the derived value is very different from the true value, e.g., $z \sim 0.7$ instead of $z \sim 2.2$. PDFs can help to minimize these because in many cases the PDF for such an object will contain two or more peaks in the redshift probability function.

Previous work on PDFs for photometric redshifts has concentrated on their derivation using a color-redshift relation, derived either empirically or from spectroscopic templates. Examples in which PDFs or χ^2 distributions for objects are shown include Lanzetta et al. (1996); Fernández-Soto et al. (1999); Kodama et al. (1999); Benítez (2000); Bolzonella et al. (2000); Firth et al. (2003), and Brodwin et al. (2006) for galaxies, and Budavári et al. (2001); Richards et al.

Electronic address: nball@astro.uiuc.edu
[1] Department of Astronomy, MC-221, University of Illinois, 1002 W. Green Street, Urbana, IL 61801, USA
[2] National Center for Supercomputing Applications, MC-257, 1205 W. Clark St, Urbana, IL 61801, USA
[3] Department of Physics, MC-704, University of Illinois, 1110 W. Green Street, Urbana, IL 61801, USA

(2001); Weinstein et al. (2004), and Wu et al. (2004) for quasars.

In this paper, we utilize objects with spectra from the Fifth Data Release (DR5; Adelman-McCarthy et al. 2007) of the Sloan Digital Sky Survey (SDSS; York et al. 2000) to train a nearest-neighbor instance-based machine learning algorithm, and perform blind tests to assess the utility of the method in assigning PDFs. We present results for main sample galaxies (MSGs; Strauss et al. 2002), luminous red galaxies (LRGs; Eisenstein et al. 2001), and quasars (Richards et al. 2002). Each of these samples have successively lower sample densities, but probe larger cosmic volumes. With our approach, it is also possible to generate PDFs for the entire SDSS photometric database. Similar work was carried out for classification probabilities by Ball et al. (2006) for the quantity P(star,galaxy,neither-star-nor-galaxy). However, such an effort is beyond the scope of the current paper, as we work solely with objects for which spectra are available.

In addition to the SDSS, we cross-match the SDSS data to the Third Data Release (GR3; Morrissey et al. 2007) of the Galaxy Evolution Explorer All-Sky Imaging Survey (GALEX AIS; Martin et al. 2005). This provides an additional two bands in the near- and far-UV, giving useful information by extending the SED coverage, e.g., so that at $z < 0.3$ we can potentially sample information for quasars from both the Mg II line in the UV and Hα in the optical.

2. DATA

We utilize data from the SDSS DR5 and the GALEX AIS GR3. In the SDSS, we select primary non-repeat observations of objects with spectra classified as galaxies and quasars (`specClass` = `galaxy`, `qso` or `hiz_qso`) in the `specObj` view of the Catalog Archive Server (CAS). In GALEX we select photometric objects with `primary_flag` = 1 from its similar CAS interface. All object attributes and errors used are from these sources. The data are retrieved via SQL queries.

In the SDSS, the object attributes retrieved are the magnitudes, *ugriz*, the associated errors derived from photon statistics (Stoughton et al. 2002), and the spectral type. The SDSS imaging covers the wavelength range $3000\text{Å} - 10,500\text{Å}$, and the spectra $3800\text{Å} - 9200\text{Å}$. Each magnitude is measured in four different apertures: *PSF*, *fiber*, *Petrosian*, and *model*; and we require all magnitudes to be within the range $0 < mag < 40$, and magnitude errors to be within $0 < magErr < 10$. Much tighter cuts could reasonably be applied; but we simply wish to eliminate extreme outlying values that are entirely unphysical (e.g., -9999) as they can cause instability in the learning algorithm. We also note that less outlying values should be easily accounted for by the learning process. Throughout this work, the SDSS magnitudes are corrected for Galactic extinction using the dust maps of Schlegel et al. (1998), and the GALEX magnitudes are corrected using the $B - V$ (`e_bv`) term inferred from these maps using the standard formula of Cardelli et al. (1989).

We subdivide the objects into three samples: main sample galaxies, luminous red galaxies, and quasars. Each sample is subject to additional cuts as appropriate. For MSGs (`specClass` = 2), we require, following Strauss et al. (2002), Petrosian magnitude $r < 17.77$, `zWarning` = 0, `zStatus` > 2, and `zConf` > 0.85. For LRGs, we apply the selection criteria of Eisenstein et al. (2001), resulting in `specClass` = 2, `primTarget` = `TARGET_GALAXY_RED`, $z > 0.2$, `zWarning` = 0, `zStatus` > 2, and `zConf` > 0.85. For quasars, we require `specClass` = 3 or 4, `zWarning` = 0, and `zStatus` > 2. We remind the reader that extinction-corrected magnitudes are used throughout. The resulting numbers of objects are 413,361 MSGs, 66,268 LRGs, and 55,743 quasars. The quasar sample is the same as that of Ball et al. (2007, hereafter B07), with the loss of three objects due to our additional restriction on the magnitude error.

The SDSS samples are cross-matched to the primary photometric objects in the `photoObjAll` view of the GALEX AIS GR3 database, using an RA+decl. tolerance (a distance on the sky) of 4 arcsec. This adds the near-UV ($1750\text{Å} - 2750\text{Å}$) and far-UV ($1350\text{Å} - 1750\text{Å}$) bands. We require the match to be unambiguous, in the sense that no SDSS object is within 4 arcsec of more than one GALEX object. Figure 1 shows histograms of the object separations. The majority of these are much smaller than 4 arcsec, indicating that our tolerance is reasonable. For the GALEX photometry, we construct samples requiring a detection in both the near and far-UV bands (`band` = 3), and a second set of samples just requiring detection in the near-UV band (`band` = 1). The latter are constructed because the samples are considerably larger, but still incorporate some UV information. We again require magnitudes in the range 0–40, and the flags `fuv_artifact` and `nuv_artifact` are required to be 0. The resulting matches consist of 59,845, 256, and 10,328 objects for MSGs, LRGs, and quasars respectively in near- and far-UV; and 100,826, 2316, and 17,110 objects for near-UV only. Several qualitative features of these matches are as expected: (1) the near-UV-only samples are larger because they only require one detection not two; (2) there are very few matches to LRGs in GALEX, because LRGs are dominated by massive red early-type galaxies with little ongoing star formation; (3) the size of the quasar matched far- and near-UV sample increases in the same proportion to the size of the GALEX dataset as a whole compared to that obtained for GALEX GR2 by B07; and (4) there are few quasar matches beyond $z \sim 2$, because the Lyman limit at 912Å has redshifted out of the GALEX bands.

Following B07, in addition to the SDSS and SDSS+GALEX datasets, we also analyze the SDSS+GALEX sample of objects, but using only SDSS features. As in B07, these datasets are referred to as *GALEX-SDSS-only*, and they enable us to quantify the level of improvement in SDSS+GALEX seen from the addition of the GALEX UV features, as opposed to possible improvement due to the sample only containing luminous quasars that appear in both SDSS and GALEX.

3. ALGORITHMS

We apply the NN instance-based learning to each of the datasets. Full details of the NN method and its extension to k-nearest neighbors (kNN) are given in B07, or in, e.g., Aha et al. (1991); Witten & Frank (2000), or Hastie et al. (2001). Briefly, the method requires a set

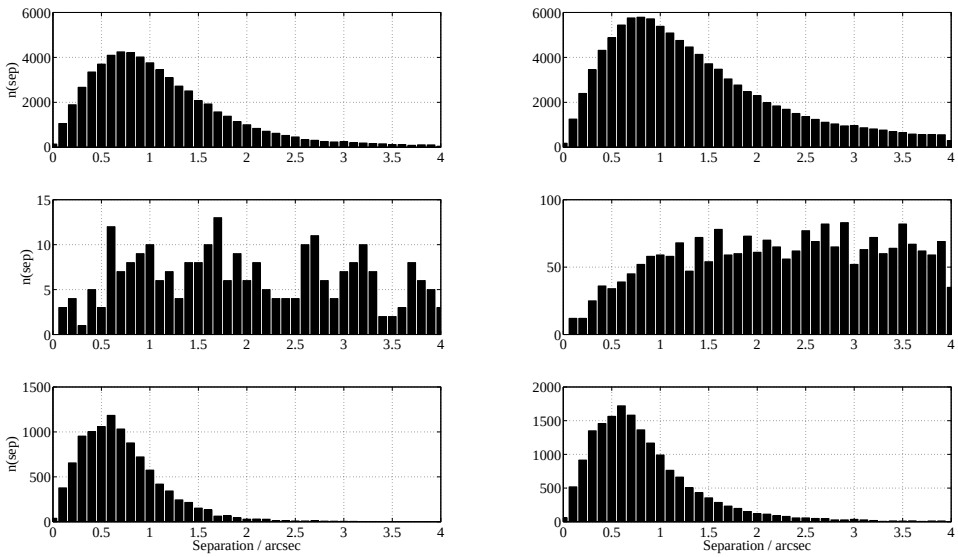

FIG. 1.— Histograms of separations in SDSS-GALEX cross-matches. The rows show main sample galaxies, luminous red galaxies, and quasars respectively. The left-hand column shows FUV+NUV, the right-hand column shows NUV-only.

of training features for each object and a target property. The algorithm then compares the position in feature space of each new object in the testing set to the training set, and assigns the target property of the nearest training set object. This may be generalized to a weighted sum of nearest neighbors, i.e., $k > 1$. The method is computationally intensive[4], but we are able to exploit its full power by utilizing nationally allocated, peer-reviewed time on the Xeon Linux cluster Tungsten at the National Center for Supercomputing Applications (NCSA), and the Java environment Data-to-Knowledge (Welge et al. 1999).

Throughout this paper, for the SDSS data the training features are the 4 colors $u-g$, $g-r$, $r-i$, and $i-z$ in the four magnitude types *PSF*, *fiber*, *Petrosian*, and *model*. This results in 16 training features. In B07, genetic algorithms were used to investigate subsets of these parameters in a systematic way; however, no subset was found that resulted in significant improvement, and indeed many subsets were worse. Preprocessing the training features with principal component analysis may remove some redundancy and save computational time, but given the aforementioned B07 result, we elected to simply use the full 16 colors throughout, in the spirit of using the full information available.

The target property is the spectroscopic redshift, which we regard as being correct as any error on this value is expected to be small compared to the photometric redshift error. When cross-matched to the GALEX data, the addition of the far-UV (FUV) and near-UV (NUV) bands gives the additional colors FUV-NUV, and four instances of NUV-u, resulting in 21 training features. The GALEX-SDSS-only sets contain the same objects as the SDSS+GALEX, but with just the assigned 16 SDSS features used.

As in B07, we standardize the training features such that each has a mean of 0 and a variance of 1. We test the performance of the algorithm by splitting each dataset into a training set, consisting of an 80% random subsample of the data, and a blind test set, consisting of the remaining 20%. The blind test set does not overlap with the training set, and this represents a realistic measure of the performance of the algorithm on unseen data within the same color regime. We perform 10-fold repeated holdout validation on each dataset, i.e., 10 different training:blind test splits, and quote the mean and standard deviation of the results for each.

The NN method as described, produces a single scalar-valued photometric redshift for each object (for each training:blind split). Therefore, to generate a probability density function in redshift for each object, we *perturb* the values of the training and blind features according to the given error on each feature. We assume the errors are Gaussian, and the perturbation is applied to the magnitude before the color is derived. This assumes that there is negligible covariance between the magnitudes. Scranton et al. (2005) show that in fact this is not necessarily the case with the available SDSS magnitude errors. However, these errors are underestimated, which cancels the covariance at the 10-20% level. We therefore use the values supplied in the SDSS DR5 for the work presented here. For each MSG we apply 10 perturbations to the features in the training set and 10 to the blind test set, giving 100 photometric redshift values per galaxy. For LRGs and quasars, we similarly apply 32 perturbations, giving 1024 values per object. The same numbers of perturbations are used for SDSS, SDSS+GALEX, and GALEX-SDSS-only.

We analyze the PDFs in terms of peaks in the probability. We establish these peaks by fitting a piecewise polynomial spline to the binned redshift counts. Peak redshifts are defined as the value at which half the integrated area lies under a peak. We set a threshold under which the area does not count, so that very low peaks are not identified. The threshold level is that which the

[4] For example, to generate the MSG PDFs requires 172048 s on 100 nodes for $n_{\rm PDF}\ n_{\rm galaxy}\ n_{\rm validation} = 100 \times 82672 \times 10$ galaxies, giving 4.8 galaxies s^{-1}.

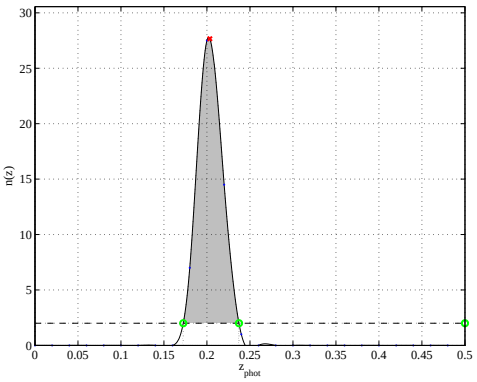

FIG. 2.— Example PDF for an SDSS DR5 main sample galaxy. The black line is the PDF, a spline fit to the binned redshifts for each object (100 for main sample galaxies, 1024 otherwise); the red cross is the peak redshift, corresponding to half of the peak area; the blue dots are the binned individual redshifts; the horizontal dashed line is the peak threshold, and the shaded areas are the PDF peaks.

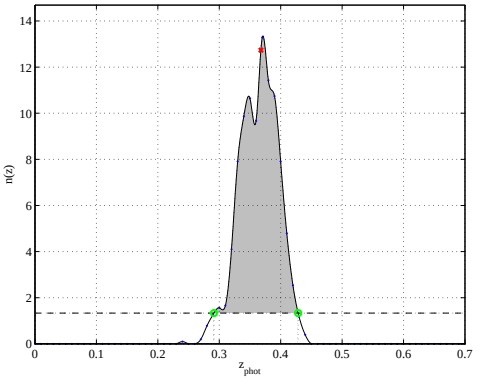

FIG. 3.— As Figure 2 but for a luminous red galaxy.

PDF would be at if it were completely flat. Thus peaks above this represent excess probability, and an object is essentially guaranteed to have at least one redshift peak. The area under a peak but also under the threshold is not included in the integral, so that the redshift values of the peaks are not pulled towards the peak centers by probability that would not otherwise count. Figures 2-5 show typical examples of object PDFs.

4. RESULTS

In general, we find the expected result that, for the SDSS, main galaxies and luminous red galaxies show mainly Gaussian-like PDFs, and quasars give a higher incidence of catastrophic failures. The addition of GALEX data does not greatly affect the galaxies, but substantially improves the results for quasars. We therefore present our galaxy results first, and then concentrate on the quasar results, for which we perform additional tests.

The $z_{\rm phot}$ versus $z_{\rm spec}$ dispersions and percentages of objects within $\Delta z < 0.1, 0.2,$ and 0.3 are tabulated for the whole blind test samples in Table 1, and for objects with a single PDF peak in Table 1. Note that the dispersions are given as σ throughout and not $\sigma/(1+z)$.

4.1. *Galaxies*

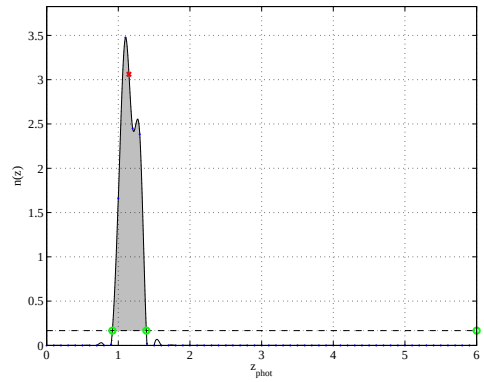

FIG. 4.— As Figure 2 but for a quasar with one peak redshift.

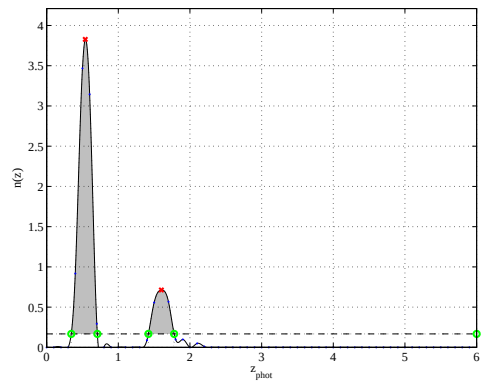

FIG. 5.— As Figure 2 but for a quasar with multiple peak redshifts.

Figure 6 shows the photometric redshift, $z_{\rm phot}$, versus spectroscopic redshift, $z_{\rm spec}$, for the 82,672 SDSS DR5 MSGs in the blind testing set. For each object shown, the value of $z_{\rm phot}$ is the mean of the set of individual redshift values that make up its PDF. The plot shows only one instance of the training:blind split from the ten used in the holdout validation, as each split divides the sample into different subsets. Also, uniquely among the datasets, the SDSS MSGs produced a small fraction (0.2%) of objects that could not be fit by the spline. These are excluded from the plot. We do not expect that the missing objects would have any significant impact on the results quoted here because (1) the raw PDFs were examined visually for these objects and did not appear unusual; and (2) earlier results based on direct examination of the PDF histogram included these missing objects and did not yield a significantly different value of σ.

For DR5 MSGs, we find that the overall RMS dispersion between $z_{\rm phot}$ and $z_{\rm spec}$, taking into account the holdout validation, is $\sigma = 0.0207 \pm 0.0001$. This is very similar to numerous previous published results, who all obtain $0.02 \lesssim \sigma \lesssim 0.025$, e.g., Brunner et al. (1997) from Galactic fields at high latitude, and in the SDSS, Tagliaferri et al. (2002); Csabai et al. (2003); Firth et al. (2003); Ball et al. (2004); Collister & Lahav (2004); Vanzella et al. (2004); Wadadekar (2005); Way & Srivastava (2006); D'Abrusco et al. (2007); Kurtz et al. (2007); Li et al. (2007); Oyaizu et al. (2008); Wang et al. (2007, 2008), and Wray & Gunn

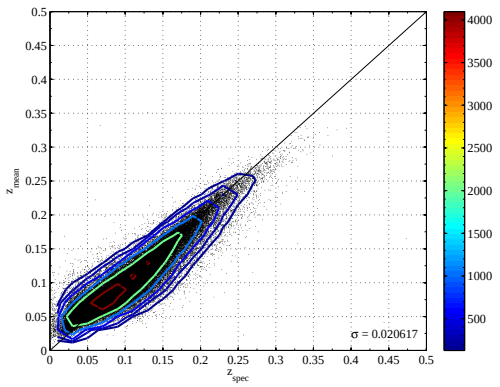

FIG. 6.— Photometric versus spectroscopic redshift for the 82,672 SDSS DR5 main sample galaxies of the blind testing set (20% of the sample). $z_{\rm phot}$ is the mean photoz from the PDF for each object. The result from a single split (of the ten used for validation) of the data into training and blind testing data is shown. σ is the RMS dispersion between $z_{\rm phot}$ and $z_{\rm spec}$.

(2007). Other methods for selecting a redshift from the PDFs, for example, the mode, median, and the same values using the binned data rather than the raw redshifts, give very similar results.

The addition of GALEX data reduces the blind testing sample size to 11,969 for FUV+NUV and 20,165 for FUV-only. For NUV-only, the dispersion is the same as the optical ($\sigma = 0.0209 \pm 0.0002$), and for FUV+NUV it is slightly worse, at $\sigma = 0.0231 \pm 0.0004$. The GALEX-SDSS-only values also show very similar dispersion, indicating that GALEX photometry is making little difference to the spread. In fact, it is not surprising that the addition of the GALEX bands does little to help, for both MSG and LRG, because the dominant source of redshift information in color space, the 4000Å break, is always redwards of the GALEX bands.

If one analyzes the $z_{\rm phot}$ values from a single run of the NN algorithm, without perturbing the input features and taking the mean, then the dispersion is higher than that derived from the mean of the PDF, typically $\sigma \sim 0.03$. However, the galaxies become much more symmetrically distributed about the $z_{\rm phot} = z_{\rm spec}$ locus. We discuss possible reasons for this in §5.1.

In the full results, if one selects galaxies with a single PDF peak, the σ value improves slightly to to $\sigma = 0.0198 \pm 0.0001$, but the values for multiple peaks are significantly higher. For galaxies with multiple peaks, if one artificially selects the peak nearest to $z_{\rm spec}$, regardless of its relative height compared to the other peaks (i.e., the 'best peak', an estimate of the best possible photoz prediction), the resulting dispersions remain similar to the single-peak value for both SDSS and SDSS+GALEX.

Given that the dispersion for the whole sample is $\sigma \sim 0.02$, consistent with numerous previous results in the literature, and the generally Gaussian nature of the PDFs, and the lack of improvement from either single peaks or the artificial best peaks, we conclude that the PDFs for MSGs are approximately optimal, given the data.

Figure 7 shows the analogous results to Figure 6 for LRGs. Again the σ value is similar to previous work, e.g., Padmanabhan et al. (2005); Collister et al. (2007); D'Abrusco et al. (2007), and Lopes (2007); and the single

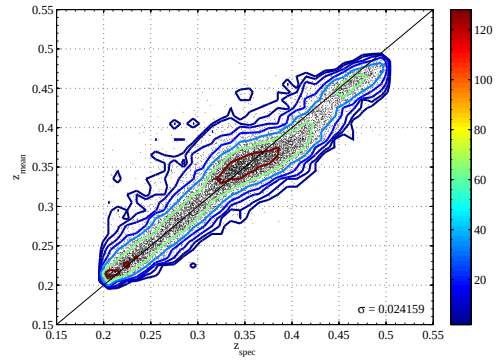

FIG. 7.— As Figure 6, but for 13,254 SDSS DR5 luminous red galaxies.

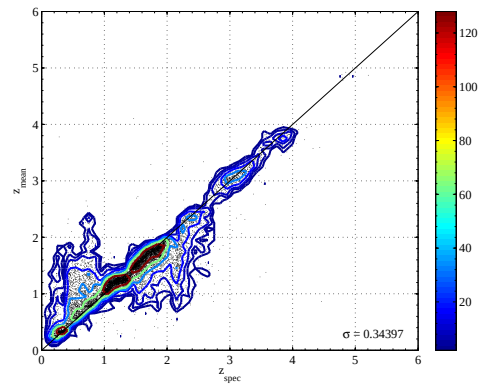

FIG. 8.— As Figure 6, but for 11,149 SDSS DR5 quasars.

and 'best' peaks again make little difference. There is a kink in the $z_{\rm phot}$ versus $z_{\rm spec}$ plot at $z \sim 0.35$, likely due to the movement of the 4000Å break between filters (see §5.1).

4.2. *Quasars*

4.2.1. *SDSS DR5*

Figure 8 shows the mean PDF $z_{\rm phot}$ versus $z_{\rm spec}$ for 11,149 blind test quasars, in a similar manner to Figures 6 and 7. The result is similar to that obtained by B07 using our multiple-nearest neighbor approach. We obtained a value of $\sigma = 0.35$ (there quoted as $\sigma^2 = 0.123 \pm 0.002$), compared to $\sigma = 0.343 \pm 0.005$ here.

Here, we are able to improve on this using the information available in the PDFs. Figure 9 shows the same data as Figure 8, but for those objects which have a single PDF peak. Although, averaged over the ten holdout validation runs, this reduces the sample size from 11,149 to 4339 ± 24, and alters the selection function, the improvement is dramatic. The dispersion is improved from $\sigma = 0.343 \pm 0.005$ to $\sigma = 0.117 \pm 0.001$. The percentage of quasars within $\Delta z < 0.1, 0.2,$ and 0.3 is increased from $53.8 \pm 0.4\%$, $72.4 \pm 0.3\%$, and $79.8 \pm 0.3\%$ [5] to $73.6 \pm 0.6\%$, $96.3 \pm 0.1\%$, and $99.3 \pm 0.1\%$. Just 33 ± 4 objects from 4339 ± 24 (0.7%) remain as catastrophics. Weinstein et al. (2004) obtained 83% of quasars within

[5] Ball et al. (2007) found $54.9 \pm 0.7\%$, $73.3 \pm 0.6\%$, and $80.7 \pm 0.3\%$. This is consistent within the errors.

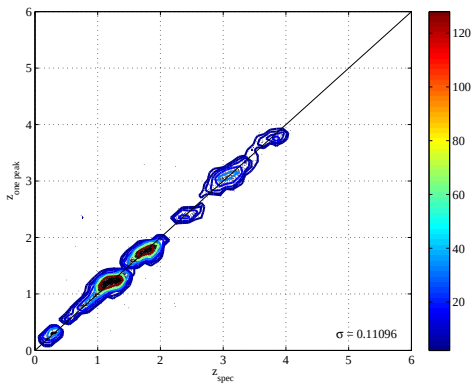

FIG. 9.— As Figure 6, but for SDSS DR5 quasars with single PDF peaks. Over the ten validation runs, the number of objects with one peak from the blind test sample of 11,149 is 4339 ± 24. The alteration of the selection function, $n(z)$, is clear, but so is the dramatic improvement in the dispersion of the remaining objects. 99.3% are within $\Delta z = 0.3$.

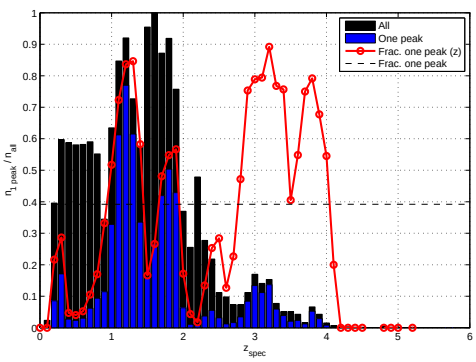

FIG. 10.— Alteration in the selection function for the subsample of SDSS DR5 quasars with one peak compared to the full sample. The horizontal dashed line shows the overall fraction of quasars with single-peaked PDFs, to which the red line would correspond if there were no alteration.

$|z_{\rm spec} - z_{\rm phot}| < 0.3$ for their whole sample, but their dispersion is much higher (cf. their figure 4 and Figure 8).

Figure 10 shows the alteration in the selection function if we insist on only one peak in the PDF. The fraction of objects with one peak is either significantly decreased or increased from the average. Significant deficits occur at $0 \lesssim z \lesssim 1$ and $1.9 \lesssim z \lesssim 2.8$, with excess in the remaining redshift ranges. These ranges correspond to the an increased dispersion of $\sigma \sim 0.5$. We discuss possible reasons why these redshift ranges are poor in §5.2.

4.2.2. *SDSS DR5 + GALEX GR3*

Figure 11 shows $z_{\rm phot}$ versus $z_{\rm spec}$ for the mean of the PDF in a similar manner to Figures 6, 7, and 8. The sample size is reduced from 11,149 to 2066, but, as shown in B07, the addition of the two GALEX bands (FUV and NUV) substantially improves the results for the remaining objects. Here, the dispersion is reduced from the SDSS value of $\sigma = 0.343 \pm 0.005$ to 0.234 ± 0.011, and the percentage of non-catastrophics is increased from $79.8 \pm 0.3\%$ to $90.8 \pm 0.5\%$. This improvement is achieved without requiring a single peak in the PDF. When this requirement is made, the sample size is 1093 ± 24, the dispersion is further improved to $\sigma = 0.106 \pm 0.016$, and

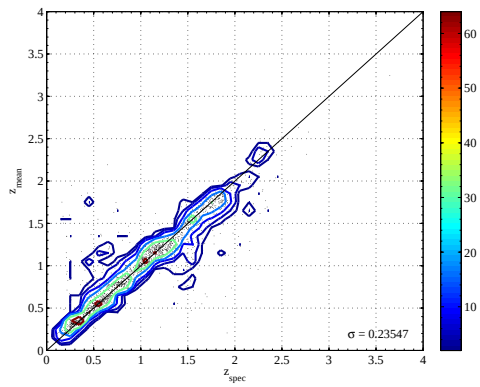

FIG. 11.— Photometric versus spectroscopic redshift for 2066 SDSS DR5 quasars incorporating near- and far-UV photometry from matching to GALEX GR3. The improvement over SDSS alone is achieved without requiring single-peaked PDFs.

$99.5 \pm 0.2\%$ of the quasars are within $\Delta z < 0.3$.

We also find that, in addition to this FUV+NUV match, the NUV-only match produces results which are almost as good, and for a sample that is 70% larger, at 3422 objects. The dispersion and percent non-catastrophics are $\sigma = 0.242 \pm 0.009$ and $90.8 \pm 0.4\%$. The results for the 1864 single-peaked objects are $\sigma = 0.109 \pm 0.010$ and $99.4 \pm 0.1\%$. This is a similar fraction of objects with a single peak as seen in the SDSS sample, with similar statistics.

An analysis of the GALEX-SDSS-only data confirm that the improvement, as in B07, is genuinely due to the addition of the GALEX data and not simply a consequence of requiring the objects to be detected by GALEX. The full results are given in Table 1.

5. DISCUSSION

5.1. *Galaxies*

For galaxies, the results are similar to previous work, as described in §4.1, except that we are now presenting PDFs for each galaxy in the form of 100 (for MSG) and 1024 (for LRG) photometric redshift estimates. These PDFs can be used to improve scientific analyses of large galaxy samples.

The overall redshifts from taking the mean values of the PDFs show similar trends to previous results, with few catastrophic failures and a smoothly decreasing incidence of objects at increasing $|z_{\rm spec} - z_{\rm phot}|$. The shapes of the PDFs generally appear Gaussian, with few widely-spaced peaks. As in previous work, Figure 6 shows that for MSGs the mean values of $z_{\rm phot}$ have a slight bias toward high values at $z_{\rm spec} \lesssim 0.1$. The single values of $z_{\rm phot}$ from the unperturbed sample do not suffer from this and are symmetrically distributed about the $z_{\rm phot} = z_{\rm spec}$ locus, but the RMS dispersion is higher, at $\sigma = 0.0284 \pm 0.0002$.

Similar behavior with respect to symmetry is seen by Ball et al. (2004) (their figure 1), who show the same biases in morphological classification, which is largely driven by the inverse concentration index. Such bias is not likely due to the lower end of the scale being zero: Ball (2004) (§3.6.3) showed that it is still present when the targets are numerically shifted, i.e., that it is the end of the scale that matters not its numerical value. A more likely explanation is that taking the mean is anal-

ogous to using multiple neighbors in the training set. Ball et al. (2004) found that when the training set for the *eClass* eigenclass spectral type (Connolly et al. 1995; Connolly & Szalay 1999; Yip et al. 2004) is cut so that it is not dominated numerically by galaxies of the eClass corresponding to early types, the resulting classifications were spread more symmetrically about the target locus, especially in the region of early types. Vanzella et al. (2004) (their figure 16) show a similar example with SDSS DR1 galaxies. Likewise, we find here when using a single neighbor that the types are more symmetrically distributed, albeit with higher dispersion. Thus it is possible that using multiple neighbors is subjecting the results to the inevitable uneven distribution of objects in parameter space, in this case colors, causing the same effect as seen in Ball et al. (2004). It appears to be a generic feature of single nearest-neighbor models (also seen for quasars) that the single neighbor produces a roughly symmetrical distribution about $z_{\rm phot} = z_{\rm spec}$, but using multiple neighbors (as in B07), or, in this case, taking the mean of the PDF, reduces the dispersion but introduces structure into the relation. However, in the latter case the structure is not usually large compared to the dispersion.

The LRGs, being constrained to be in the redshift range $0.2 < z_{\rm spec} < 0.5$ appear to suffer less from this type of bias at low and high redshift, but instead show a kink at $z_{\rm spec} \sim 0.35$. This is likely due to the 4000Å break passing between the SDSS g and r bands. Again, the single photoz values do not show this bias, but the dispersion is higher at $\sigma = 0.0318 \pm 0.0002$.

5.2. *Quasars*

As described in §4, there is a higher incidence of catastrophic failures among quasars compared to galaxies, which results in more PDFs with widely spaced peaks. Possible contributory factors include reddening, contamination of the quasar light by a host galaxy, degeneracy in the color-redshift relation, low equivalent-width lines, and unusual spectral slopes. But, in particular, bright spectral lines dropping between filters, or simulating other lines at a different redshift, are a major contributor. In Figure 12, we overplot, on the $z_{\rm phot}$ versus $z_{\rm spec}$ plot, the redshifts at which the five brightest emission lines of the composite quasar spectrum of Vanden Berk et al. (2001) cross the SDSS filter edges. The lines are, in order of flux: Lyα, C IV, C III, Mg II, and Hα. There is a very clear correspondence between the redshifts at which the emission lines cross the filters, with particularly striking examples for Mg II at $z \simeq 0.4$, Hα at $z \simeq 0.25$, and Lyα at $z \simeq 2.2$. The lower right panel overplots the five lines, showing that there is no visually significant structure that does not correspond to one of these lines. The general pattern is that in between the lines, the redshifts are less spread out, and they then jump to a new value where the lines cross. It is also noticeable that most of these discontinuities not only correspond to a line crossing a filter, but to several lines doing so in a small redshift range. All this shows that objects moving between filters, and the resulting missing information or degeneracy, is a likely cause of much of the remaining error in optical quasar photometric redshifts. This information could be used to develop 'optimal' filter sets for quasar photometric redshift estimation with future surveys.

For multiply-peaked PDFs, when the mean $z_{\rm phot}$ or $z_{\rm phot}$ of the highest peak is not correct, one of the other peaks is often close to the correct redshift. It turns out that if one artificially selects these peaks, then the results are no better than those for single peaked objects (§4.1 and §4.2.1). Nevertheless, given that these objects are a smaller subset of the full sample with a different selection function, we attempted to improve the probability of the correct peak being chosen for the whole sample by applying known prior information to the derived PDFs. This was in the form of the known redshift and magnitude distributions of quasars, and their luminosity function constructed using the values of Richards et al. (2006). Given the known $n(z)$ of spectroscopic quasars, the photometric redshift is more likely to be in a region of high $n(z)$, thus weighting the PDF accordingly may increase the incidence of the highest peak being the correct one. Similarly, a quasar may be assigned a redshift that, given its apparent magnitude, would make it unrealistically faint or luminous. This can be downweighted by applying the magnitude distribution or luminosity function. However, we found that the prior information did not improve the redshift statistics. This is not especially surprising because an empirical training algorithm such as the one we are using implicitly takes into account the priors by its use of the training set. Thus further weighting will not add as much information as it would, for example, in a template-based method.

Finally, we investigated the effect of altering the threshold (§3) above which the area under the PDF is counted as part of the redshift peaks. The main effect is to alter the relative quality of the subset of objects with single peaks. A higher threshold value will produce more objects with one peak but the sample is lower quality, and vice-versa. The threshold quoted is chosen because it corresponds to what a flat PDF would be, but it also appears approximately optimal when the percentage of objects with one peak is compared to their dispersion or Δz. For much higher thresholds, σ increases relatively faster than the sample size; for much lower thresholds, the dispersion does not decrease as fast as the sample size.

From the investigations presented in this section, we conclude that, for SDSS DR5 optical data, it is unlikely that the MSG, LRG, or quasar redshifts can be significantly improved without the addition of new data. However, we have shown that, for quasars, if one is willing to discard the percentage of the objects that have more than one PDF peak, the photometric redshifts are significantly improved.

6. CONCLUSIONS

We apply nearest neighbor machine learning to objects with spectra in the Sloan Digital Sky Survey Data Release 5 (SDSS DR5). We subdivide the objects into 413,361 main sample galaxies, 66,268 luminous red galaxies, and 55,743 quasars, both in the SDSS, and matched to the Galaxy Evolution Explorer All Sky Imaging Survey Third Data Release (GALEX AIS GR3). Each sample is divided into a training set consisting of 80% of the objects, and a blind testing set of the remaining 20%. The algorithm assigns a full probability density function (PDF) in photometric redshift to each object in

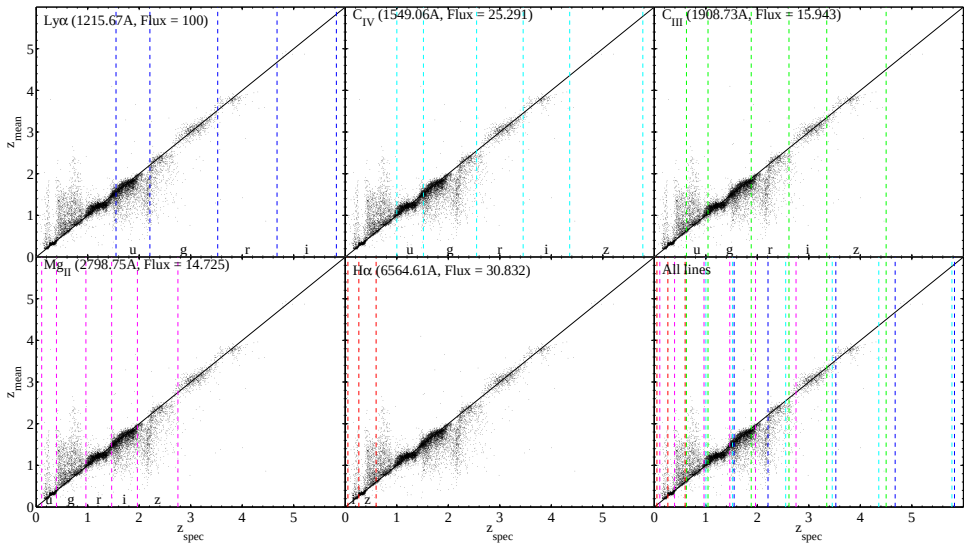

FIG. 12.— Redshifted filters overplotted on $z_{\rm phot}$ versus $z_{\rm spec}$ for SDSS DR5 quasars for the five brightest emission lines. There is a clear correspondence between the redshifts at which the emission lines cross SDSS filters and the occurrence of structure in the plot. The bottom right-hand panel shows all five lines superimposed. Visually, there is no significant structure that does not correspond to one of these lines, and often several lines are close in redshift.

the blind testing set by perturbing the input features describing the objects, in this case the colors, according to the magnitude errors. For main sample galaxies (MSGs), each PDF is formed from 100 photometric redshift values, and for luminous red galaxies (LRGs) and quasars, 1024 values.

We use the spectroscopic redshifts to test the utility of the method and find that the RMS dispersions between the photometric and spectroscopic redshifts are $\sigma = 0.0207 \pm 0.0001$, $\sigma = 0.0243 \pm 0.0002$, and $\sigma = 0.343 \pm 0.005$ for MSGs, LRGs, and quasars, respectively. The quoted errors are generated from ten-fold repeated holdout validation in the form of ten different training-to-blind-testing set splits of the data. Galaxy values are similar to previous studies, and quasar values are consistent with Ball et al. (2007). Cross-matching to GALEX reduces the dispersion for quasars to $\sigma = 0.234 \pm 0.011$ for the 10,328 matching objects. The improvement is due to the GALEX photometry and not simply an artifact of requiring the objects to be detected by GALEX. It may be possible to improve the galaxy results by incorporating morphological information into the training set, for example the inverse concentration index, but the improvement would likely be small for these data.

For quasars, use of the PDFs enables us to construct subsamples which show dramatically improved statistics. In particular, selection of objects with a single PDF peak in the full SDSS reduces the sample size by two thirds but improves the dispersion from $\sigma = 0.343 \pm 0.005$ to $\sigma = 0.117 \pm 0.010$, with a substantial increase in the number of non-catastrophic failures (photometric minus spectroscopic redshift less than 0.3) from $79.8 \pm 0.3\%$ to $99.3 \pm 0.1\%$. The equivalent statistics for the GALEX sample are $\sigma = 0.234 \pm 0.011$ to $\sigma = 0.106 \pm 0.016$, and $90.8 \pm 0.5\%$ to $99.5 \pm 0.2\%$. The improved samples alter the selection function, but there is a good fraction of high percentage one-peak-regions over almost the whole redshift range.

We attempted weighting the PDFs according to known prior information on the distributions of quasars in redshift, apparent magnitude, absolute magnitude from the photometric redshift, and absolute magnitude from the luminosity function of Richards et al. (2006), but this did not improve the statistics. We also derived statistics for the PDF peak closest to the spectroscopic redshift regardless of its height, and found that these do not improve on the statistics for quasars with one peak. We overplot the redshifts at which bright quasar emission lines cross the SDSS filter edges and find that there is a clear correspondence with changes in the photometric redshift dispersion. This strongly suggests that a large fraction of the remaining poor redshifts are caused by lines disappearing at filter edges, or simulating other lines. We conclude that further improvement requires better data in the form of spectra for fainter objects, or a larger number of filters, both within and beyond the optical range (in particular the UV and IR).

The NN method is conceptually simple, and, once the dataset has been selected, has no adjustable parameters. This means that, unlike most machine learning algorithms, all of the information in the training data is used, and all of the computation contributes to the final result, rather than exploring parameter space and generating mostly unused results.

Future work includes the application of the methods here to more and deeper optical data, e.g., the full SDSS photometric database, the 2QZ, 2SLAQ, SDSS Southern, VVDS, DEEP2, and COSMOS surveys, and the addition of infrared data via UKIDSS and S-COSMOS. For quasars, a further useful addition would be a filter set that is customized for quasars rather than the stars and galaxies typical of optical surveys, or more and narrower filters, as in COMBO-17 (e.g., Wolf et al. 2004), but for a wider field. The filters should overlap to minimize errors from line movement across filter edges.

We thank the referee for a prompt and useful report.

The authors acknowledge support from NASA through grants NN6066H156 and NNG06GF89G, from Microsoft Research, and from the University of Illinois. The authors made extensive use of the storage and computing facilities at the National Center for Supercomputing Applications and thank the technical staff for their assistance in enabling this work.

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is http://www.sdss.org/.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, Cambridge University, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPA), the Max-Planck-Institute for Astrophysics (MPIA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

Based on observations made with the NASA Galaxy Evolution Explorer. GALEX is operated for NASA by the California Institute of Technology under NASA contract NAS5-98034.

Data To Knowledge (D2K) software, D2K modules, and/or D2K itineraries, used by us, were developed at the National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana-Champaign.

This research has made use of NASA's Astrophysics Data System.

REFERENCES

Adelman-McCarthy, J. K. et al. 2007, ApJS, 172, 634
Aha, D. W., Kibler, D., & Albert, M. K. 1991, Machine Learning, 6, 37
Ball, N. M. 2004, PhD thesis, Astronomy Centre, University of Sussex, UK
Ball, N. M., Brunner, R. J., Myers, A. D., Strand, N. E., Alberts, S., Tcheng, D., & Llorà, X. 2007, ApJ, 663, 774
Ball, N. M., Brunner, R. J., Myers, A. D., & Tcheng, D. 2006, ApJ, 650, 497
Ball, N. M., Loveday, J., Fukugita, M., Nakamura, O., Okamura, S., Brinkmann, J., & Brunner, R. J. 2004, MNRAS, 348, 1038
Benítez, N. 2000, ApJ, 536, 571
Blake, C. & Bridle, S. 2005, MNRAS, 363, 1329
Bolzonella, M., Miralles, J.-M., & Pelló, R. 2000, A&A, 363, 476
Brodwin, M. et al. 2006, ApJ, 651, 791
Brunner, R. J., Connolly, A. J., Szalay, A. S., & Bershady, M. A. 1997, ApJ, 482, L21
Budavári, T. et al. 2001, AJ, 122, 1163
Cardelli, J. A., Clayton, G. C., & Mathis, J. S. 1989, ApJ, 345, 245
Collister, A. et al. 2007, MNRAS, 375, 68
Collister, A. A. & Lahav, O. 2004, PASP, 116, 345
Connolly, A. J. & Szalay, A. S. 1999, AJ, 117, 2052
Connolly, A. J., Szalay, A. S., Bershady, M. A., Kinney, A. L., & Calzetti, D. 1995, AJ, 110, 1071
Csabai, I. et al. 2003, AJ, 125, 580
D'Abrusco, R., Staiano, A., Longo, G., Brescia, M., Paolillo, M., De Filippis, E., & Tagliaferri, R. 2007, ApJ, 663, 752
Eisenstein, D. J. et al. 2001, AJ, 122, 2267
Fernández-Soto, A., Lanzetta, K. M., & Yahil, A. 1999, ApJ, 513, 34
Firth, A. E., Lahav, O., & Somerville, R. S. 2003, MNRAS, 339, 1195
Hastie, T., Tibshirani, R., & Friedman, J. 2001, The Elements of Statistical Learning, Springer Series in Statistics (New York: Springer-Verlag)
Kodama, T., Bell, E. F., & Bower, R. G. 1999, MNRAS, 302, 152
Kron, R. G. 1995, PASP, 107, 766
Kurtz, M. J., Geller, M. J., Fabricant, D. G., Wyatt, W. F., & Dell'Antonio, I. P. 2007, AJ, 134, 1360
Lahav, O. & Suto, Y. 2004, Living Reviews in Relativity, 7, 8
Lanzetta, K. M., Yahil, A., & Fernandez-Soto, A. 1996, Nature, 381, 759
Lawrence, A. 2007, Preprint, arXiv/0704.0809
Li, L.-L., Zhang, Y.-X., Zhao, Y.-H., & Yang, D.-W. 2007, Chinese Journal of Astronomy and Astrophysics, 7, 448
Lopes, P. A. A. 2007, MNRAS, 380, 1608
Martin, D. C. et al. 2005, ApJ, 619, L1
Morrissey, P. et al. 2007, ApJS, 173, 682
Oyaizu, H., Lima, M., Cunha, C. E., Lin, H., Frieman, J., & Sheldon, E. S. 2008, ApJ, 674, 768
Padmanabhan, N. et al. 2005, MNRAS, 359, 237
Reshetnikov, V. P. 2005, Physics Uspekhi, 48, 1109
Richards, G. T. et al. 2001, AJ, 122, 1151
—. 2002, AJ, 123, 2945
—. 2006, AJ, 131, 2766
Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525
Scranton, R., Connolly, A. J., Szalay, A. S., Lupton, R. H., Johnston, D., Budavari, T., Brinkman, J., & Fukugita, M. 2005, Preprint, astro-ph/0508564
Stoughton, C. et al. 2002, AJ, 123, 485
Strauss, M. A. et al. 2002, AJ, 124, 1810
Tagliaferri, R., Longo, G., Andreon, S., Capozziello, S., Donalek, C., & Giordanoet, G. 2002, Preprint, astro-ph/0203445
Vanden Berk, D. E. et al. 2001, AJ, 122, 549
Vanzella, E. et al. 2004, A&A, 423, 761
Wadadekar, Y. 2005, PASP, 117, 79
Wang, D., Zhang, Y. X., Liu, C., & Zhao, Y. H. 2007, MNRAS, 382, 1601
Wang, D., Zhang, Y.-X., Liu, C., & Zhao, Y.-H. 2008, Chinese Journal of Astronomy and Astrophysics, 8, 119
Way, M. J. & Srivastava, A. N. 2006, ApJ, 647, 102
Weinstein, M. A. et al. 2004, ApJS, 155, 243
Welge, M., Hsu, W. H., Auvil, L. S., Redman, T. M., & Tcheng, D. 1999, in 12th National Conference on High Performance Networking and Computing (SC99)
Witten, I. H. & Frank, E. 2000, Data Mining (San Francisco: Morgan Kaufmann)
Wolf, C. et al. 2004, A&A, 421, 913
Wray, J. J. & Gunn, J. E. 2007, Preprint, arXiv/0707.3443
Wu, X.-B., Zhang, W., & Zhou, X. 2004, Chinese Journal of Astronomy and Astrophysics, 4, 17
Yip, C.-W. 2007, Preprint, arXiv/0706.4484
Yip, C. W. et al. 2004, AJ, 128, 585
York, D. G. et al. 2000, AJ, 120, 1579

TABLE 1

Photometric redshift PDF statistics. For quasars, the improvements resulting from cross-matching to the GALEX UV bands are clear. For SDSS main sample galaxies (MSGs) and luminous red Galaxies (LRGs), the cross-match produces little improvement, as expected. Quoted errors are the standard deviation from ten-fold repeated holdout validation. The Δz ($|z_{\rm spec} - z_{\rm phot}|$) thresholds are 0.01, 0.02, and 0.03 for MSGs and LRGs, and 0.1, 0.2, and 0.3 for quasars.

Dataset	Objects	Subset	$\text{SampleSize}_{\text{all}}$	RMS_{all}	$\Delta z < 0.01, 0.1$ (%)	$\Delta z < 0.02, 0.2$ (%)	$\Delta z < 0.03, 0.3$ (%)
SDSS	MSG	-	82,672	0.0207 ± 0.0001	44.8 ± 0.2	72.9 ± 0.2	87.1 ± 0.1
SDSS	LRG	-	13,254	0.0243 ± 0.0002	40.1 ± 0.3	68.2 ± 0.4	84.0 ± 0.4
SDSS	QSO	-	11,149	0.343 ± 0.005	53.8 ± 0.4	72.4 ± 0.3	79.8 ± 0.3
SDSS+GALEX	MSG	FUV+NUV	11,969	0.0231 ± 0.0004	37.0 ± 1.0	65.6 ± 0.9	83.5 ± 0.5
SDSS+GALEX	MSG	NUV only	20,165	0.0209 ± 0.0002	42.4 ± 0.2	71.5 ± 0.2	86.9 ± 0.2
SDSS+GALEX	LRG	FUV+NUV	51	0.0304 ± 0.0035	31.4 ± 4.0	57.6 ± 5.4	73.1 ± 4.5
SDSS+GALEX	LRG	NUV only	463	0.0260 ± 0.0018	38.1 ± 1.8	66.6 ± 1.3	81.4 ± 1.4
SDSS+GALEX	QSO	FUV+NUV	2066	0.234 ± 0.011	71.8 ± 0.6	86.4 ± 0.7	90.8 ± 0.5
SDSS+GALEX	QSO	NUV only	3422	0.242 ± 0.009	68.2 ± 0.7	85.4 ± 0.6	90.8 ± 0.4
GALEX-SDSS-only	MSG	FUV+NUV	11,969	0.0230 ± 0.0004	38.4 ± 0.5	67.0 ± 0.5	83.4 ± 0.5
GALEX-SDSS-only	MSG	NUV only	20,165	0.0224 ± 0.0003	39.6 ± 0.2	68.2 ± 0.3	84.2 ± 0.4
GALEX-SDSS-only	LRG	FUV+NUV	51	0.0292 ± 0.0030	29.6 ± 5.4	58.0 ± 5.6	74.3 ± 4.2
GALEX-SDSS-only	LRG	NUV only	463	0.0257 ± 0.0018	38.9 ± 2.4	66.7 ± 1.9	82.0 ± 1.2
GALEX-SDSS-only	QSO	FUV+NUV	2066	0.314 ± 0.013	64.2 ± 0.9	79.9 ± 0.5	85.5 ± 0.5
GALEX-SDSS-only	QSO	NUV only	3422	0.336 ± 0.010	57.1 ± 0.8	75.1 ± 0.8	81.9 ± 0.7

TABLE 2

As Table 1, but for the subsamples of objects with a single peak in the redshift probability density function. The improvement is particularly dramatic for the SDSS quasar sample. Unlike the full samples, the sample size now contains a confidence interval due to the holdout validation, but the variation is small, of order 1%.

Dataset	Objects	Subset	$\text{SampleSize}_{\text{1peak}}$	$\text{RMS}_{\text{1peak}}$	$\Delta z < 0.01, 0.1$ (%)	$\Delta z < 0.02, 0.2$ (%)	$\Delta z < 0.03, 0.3$ (%)
SDSS	MSG	-	$71,236 \pm 145$	0.0198 ± 0.0009	45.9 ± 0.2	74.2 ± 0.2	88.1 ± 0.1
SDSS	LRG	-	$11,231 \pm 87$	0.0223 ± 0.0002	40.4 ± 0.4	69.4 ± 0.4	85.4 ± 0.4
SDSS	QSO	-	4339 ± 24	0.117 ± 0.001	73.6 ± 0.6	96.3 ± 0.1	99.3 ± 0.1
SDSS+GALEX	MSG	FUV+NUV	7276 ± 164	0.0214 ± 0.0004	38.4 ± 1.4	67.7 ± 1.2	85.4 ± 0.6
SDSS+GALEX	MSG	NUV only	$12,077 \pm 106$	0.0191 ± 0.0002	45.4 ± 0.2	74.8 ± 0.3	89.3 ± 0.2
SDSS+GALEX	LRG	FUV+NUV	6.5 ± 2.3	0.0204 ± 0.0170	41.5 ± 19	82.5 ± 20	93.5 ± 11
SDSS+GALEX	LRG	NUV only	187.3 ± 8.6	0.0198 ± 0.0015	43.0 ± 3.3	75.2 ± 2.4	88.7 ± 2.3
SDSS+GALEX	QSO	FUV+NUV	1093 ± 24	0.106 ± 0.016	83.8 ± 0.7	97.9 ± 0.3	99.5 ± 0.2
SDSS+GALEX	QSO	NUV only	1864 ± 32	0.109 ± 0.010	80.0 ± 0.8	97.2 ± 0.3	99.4 ± 0.1
GALEX-SDSS-only	MSG	FUV+NUV	6284 ± 197	0.0211 ± 0.0003	40.5 ± 0.5	69.8 ± 0.6	85.6 ± 0.5
GALEX-SDSS-only	MSG	NUV only	$11,892 \pm 106$	0.0205 ± 0.0002	42.7 ± 0.4	71.7 ± 0.4	86.7 ± 0.4
GALEX-SDSS-only	LRG	FUV+NUV	4.6 ± 2.4	0.0169 ± 0.0075	41.9 ± 24	81.6 ± 18	89.7 ± 15
GALEX-SDSS-only	LRG	NUV only	196.4 ± 8.8	0.0206 ± 0.0017	43.6 ± 4.7	72.4 ± 4.0	87.3 ± 2.7
GALEX-SDSS-only	QSO	FUV+NUV	933 ± 20	0.143 ± 0.031	81.5 ± 1.2	97.2 ± 0.7	98.8 ± 0.4
GALEX-SDSS-only	QSO	NUV only	1542 ± 18	0.123 ± 0.015	77.1 ± 0.7	96.7 ± 0.4	99.1 ± 0.2