draft 2012-2-04 not ready for distribution
Preprint typeset using LaTeX style emulateapj v. 11/10/09

STAR–GALAXY CLASSIFICATION IN MULTI-BAND OPTICAL IMAGING

Ross Fadely[1], David W. Hogg[2,3], & Beth Willman[1]

draft 2012-2-04 not ready for distribution

ABSTRACT

Current and next–generation wide–field, ground–based optical surveys, like PanSTARRs, DES, and LSST, will survey large portions of the sky to limiting magnitudes of $r \gtrsim 24$. The problem of source classification, particularly that of separating stars from galaxies, is severe at such depths since the number of unresolved galaxies quickly overwhelm halo star counts. We investigate the advantages and shortcomings of various photometric classification techniques using COSMOS *ugriz* data, focusing on the most unresolved objects (full width half maxima $< 0.2''$). We consider two broad classes of probabilistic classification approaches, spectral energy distribution template fitting and data–driven Support Vector Machine (SVM) techniques. For template fitting, we investigate Maximum Likelihood (ML) methods and present a new Hierarchical Bayesian (HB) method, in which we learn the prior distribution of template probabilities by optimizing the likelihood for the entire dataset. SVM approaches require a set of training data to classify unknown sources. We consider an optimistic, best–case scenario (SVM$_{best}$), where the training data is a random sampling of the data in question. In addition, we consider the scenario where the SVM is trained only using the highest signal–to–noise (S/N) data (SVM$_{high\,S/N}$). We find that our HB method generally outperforms ML approaches, and delivers $\sim 80\%$ completeness in both star and galaxy samples, with purity of $\sim 40-90\%$ and $\sim 70-90\%$ for stars and galaxies, respectively, depending on the sample fraction. We using the Receiver Operating Characteristic curve to assess the relative performance of the approaches, focusing on the Area Under the Curve (AUC) statistic. We find AUC values of 0.92, 0.88, 0.84, 0.77 for SVM$_{best}$, HB, ML, and SVM$_{high\,S/N}$. We conclude, therefore, that a well trained SVM will outperform the template fitting methods examined here. However, if trained on data with high S/N, SVMs perform worse than template fitting. Thus, HB template fitting may prove to be a useful and important method for source classification in future surveys.

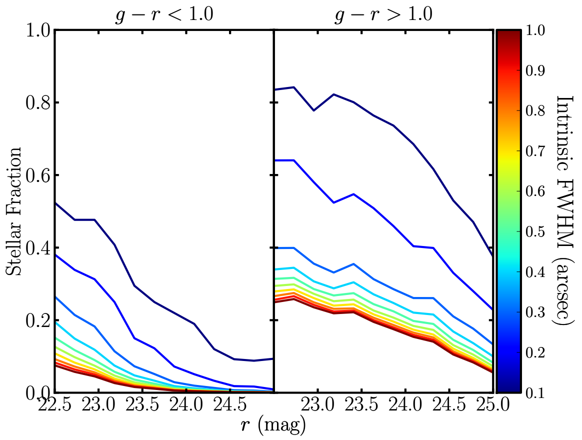

Fig. 1.— The stellar fraction of COSMOS sources as a function of magnitude, for sources with $g-r<1$ (left) and $g-r>1$ (right). Colored curves indicate the upper limit in intrinsic full–width half–maximum (FWHM) allowed in the sample. Considering sources with FWHM $< 0.2''$, red sources are only $\sim 20-60\%$ stars, while blue sources are $< 40\%$ stars even at apparent magnitudes of $r = 22.5$.

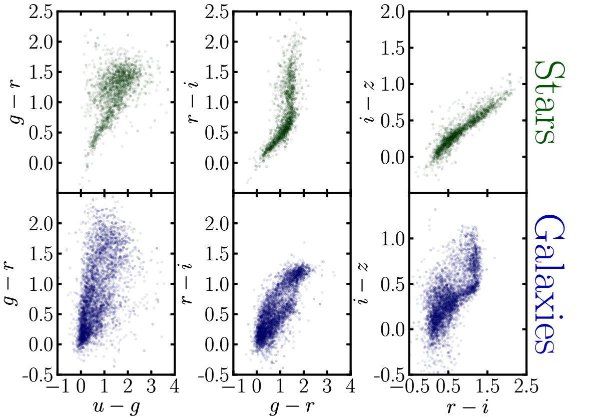

Fig. 2.— The color–color space distribution of point sources (FWHM $< 0.2''$) in the COSMOS catalog. It is clear that stars in the sample follow a tight locus in all slices of color–color space, while galaxies are more generally distributed. Even so, comparison by eye reveals significant overlap between stars and galaxies, particularly for bluer sources.

1. INTRODUCTION

2. PROBABILISTIC PHOTOMETRIC CLASSIFICATION

The goal of photometric classification of an astronomical source is to convert the input flux data $\boldsymbol{F}$ into a likelihood that the object is of a given type (considered a star S or a galaxy G here). That is, the classification algorithm produces the likelihoods $p(\boldsymbol{F}|S)$ and $p(\boldsymbol{F}|G)$ and decides classification by comparing the ratio of the likelihoods -

[1] Haverford College, Department of Physics and Astronomy, 370 Lancaster Ave., Haverford, PA, 19041
[2] Center for Cosmology and Particle Physics, Department of Physics, New York University, 4 Washington Place, New York, NY 10003, USA
[3] Max-Planck-Institut für Astronomie, Königstuhl 17, 69117 Heidelberg, Germany

$$\frac{p(\boldsymbol{F}|S)}{p(\boldsymbol{F}|G)} = \Omega \quad , \tag{1}$$

where Ω is the odds ratio which decides the classification threshold. A natural threshold is $\Omega = 1$, which may be modified to obtain more pure or complete samples, depending on the science case.

Algorithmically there are a large number of approaches which produce probabilistic classifications. Generally, these fall into i) physically based methods - those which have theoretical or empirical models for what type of physical object a source is, or ii) data driven methods - those which use real data with known classifications to construct a model for new data. Here we consider two physically based template fitting approaches, and one data driven approach. Below in this section, we present the conceptual basis for each of the three methods. In Section 3, we discuss the specific details, choices, and assumptions made for each of our classification methods. Finally, in Section 5 we show the performance of the algorithms, and discuss the advantages and limitations related to their use as classifiers.

2.1. *Template Fitting: Maximum Likelihood (ML)*

One common method for inferring spectral and classification properties from broad-band photometric data is template fitting. The principle is relatively straightforward. First the user supplies a set of spectral templates that represent the possible SED configurations of both stars S and galaxies G. Then each of the templates is convolved to with known throughput response functions in the various filters to yield a model for the data. Finally, the quality of the model is assessed against the data.

Once broad-band model values are computed for a spectral template, the template model is fully specified except for a normalization constant C. For a given data point i, value of C_i is proportional to the total luminosity of the source divided by the luminosity distance squared. This value of C_i is unknown and must be 'fit' to the data. The most common approach to determining C_i is optimizing to the value which returns the lowest χ^2, given the data. Determining the best-fitting model is known as a Maximum Likelihood (ML) approach to template fitting. After assessing the ML model for all the templates, classification is straight forward - one need only to compare the lowest star χ^2 to the lowest galaxy χ^2. In other words, $\chi^2_S - \chi^2_G = \ln(\Omega)$ is the classification criteria (cf. Equation 1).

2.2. *Template Fitting: Hierarchical Bayesian (HB)*

We present a new, Hierarchical Bayesian (HB) approach to template fitting. For our approach, we seek to capture all the information present in the data, given the measurement and model uncertainties.

While a useful baseline for template fitting, ML approaches ignore information about how likely a source is to be S or G. For instance, consider the scenario where a G model fits data $\boldsymbol{F}_i$ only *slightly* better than the best S model, while all other G models give poor fits and all other S models give nearly as likely fits. In this case, ignoring all other S models besides the best is clearly the wrong thing to do, since the data is clearly stating that S models are much more favored *generally.* Indeed, capturing this information is much more informative than simply discarding everything but the best (ML) models.

To capture this information, we seek to *marginalize* over all possibilities of what a star or a galaxy can be. That is, we want to consider the total probability that a source belongs to a certain classification (S or G). In the case of template fitting, this consists of summing up the probabilities associated with all of the S templates, as well as all the G templates (across redshift). Each of these probabilities assigned to the templates are, in turn, also marginalized likelihoods. For each fit, we compute the total likelihood of the fit by marginalizing over the uncertainty in fitting coefficient C_i. This marginalization is also easy to compute, it is simply the total probability of a Gaussian distribution with variance $\sigma^2_{C_i}$ - a value which is returned using least squares fitting techniques (e.g., Hogg et al. 2010).

By Bayes' theorem, the act marginalization requires we specify the prior probability that any object might have a given SED template (at a given redshift). These distributions might be chosen to be uninformative (e.g., flat), informed by knowledge from outside studies, or informed by the data themselves. The latter approach, referred to as a hierarchical model, is widely used in Statistics (e.g., Gelman et al. 2003) and is beginning to be used in Astronomy (Shu et al. 2012). The benefits of hierarchical approaches are many – they have been shown to increase probabilistic performance [**RF: Need cites – Hogg, any suggestions???**], and require no additional knowledge outside the given data. Functionally, this consists of parameterizing your prior distribution (e.g., the mean and variance of a Normal distribution), and varying these parameters (known as hyperparameters) to determine the probability of *all* the data under *all* the models.

For our work here, we choose to optimize the hyperparameters of the template prior distributions to return the maximum marginalized likelihood of all the data. A brief description of the functional form of these priors is given below in Section 3.2. A detailed, step–by–step description of our HB inferential procedure is provided in Appendix A and open–source `C` code is available at `http://github.com/rossfadely/star-galaxy-classification`.

2.3. *Support Vector Machine (SVM)*

A support vector machine (SVM) is a type of machine learning algorithm particularly well suited to the problem of classification. SVM algorithms are frequently used in non-astronomical problems, and are considered the gold standard for any new classification methods developed by machine learners. A SVM has recently been applied to the problem of source classification in the PanSTARRs photometric pipeline (Saglia et al. 2012), though on very high signal to noise data (SDSS DR8, $r \lesssim 19$). Here, we consider the utility and performance of SVM algorithms in a new classification regime, where the data is of lower signal to noise (described in Section 4), and the number of unresolved galaxies is comparable or more than the number of stars.

SVM algorithms work by taking a set of 'training' data with known classifications, and constructing a set of hyperplanes in a higher dimensional space that best separates the properties of each class. Here, an SVM algorithm trains on a set of data to construct the hyperplane(s) which best separate the colors and apparent magnitude[4] of stars and galaxies. For non-linear classification problems, a kernel function is used to keep computation reasonable. For more details on the SVM technique, please see [**RF: where to send people??**].

3. SPECIFICS AND IMPLEMENTATION

3.1. *ML Template Fitting*

Template based star–galaxy classification relies on the use of spectral energy distribution templates which (as best as possible) span the space of colors for both stars and galaxies. We consider a large number of stellar templates in order to compose the best possible stellar model library. First, we adopt the Pickles (1998) set of empirically derived SEDs, which span O5 to M10 stars for both main sequence, giant, and supergiant stars. The vast majority of the SEDs in the Pickles library have solar abundances, so we supplement the library with theoretical SEDs from Castelli-Kurucz (CK) (Castelli & Kurucz 2004). We use CK models with abundances ranging from $-2.5 \leq$[Fe/H]≤ 0.0, surface gravities ranging from $3.0 \leq \log(g) \leq 0.0$, and effective temperatures from $3500 \leq T_{eff} \leq 10000$ K. We also consider the possibility that many stars ought to be in binaries and combine the flux calibrated CK models of same metallicity with each other to construct binary star templates. Finally, since the majority of stars in typical datasets ought to be red, low-mass M dwarfs we also consider SDSS L0 – M9 dwarf templates provided by Bochanski et al. (2007). These templates have been extended by Bochanski et al. into the near infrared, but lack data for wavelengths shorter than 4000 Å. We extend these templates down to the 3000 Å using a main sequence CK model with T_{eff} = 3500K. Details of this extension are likely to be unimportant, since the flux of such stars between $3000 - 4000$ Å is negligible. Our final combined list of templates are 131 from the Pickles library, 255 from the CK library, 10 from Bochanski et al. (2007), and 917 binary templates constructed from the CK library, for a total of 1313 stellar templates.

We select for our galaxy templates those used by the COSMOS team, described in (Ilbert et al. 2009), provided publicly through the `Le Phare` photometric redshift package [5] (Arnouts et al. 1999; Ilbert et al. 2006). These templates consist of galaxy SEDs from Polletta et al. (2007), encompassing 7 elliptical and 12 spiral (S0-Sd) SEDs. Additionally, 12 representative starburst SEDs are included, which were added by Ilbert et al. (2009) to provide a more extensive range of blue colors. Templates from Polletta et al. (2007) include effects of dust extinction, since they were selected to fit spectral sources in the VIMOS VLT Deep Survey (Le Fèvre et al. 2005). We do not consider any additional dust extinction beyond these fiducial templates. In order to model our galaxies across cosmic time, we redshift these templates on a discrete linear grid of redshifts, ranging from 0 to 4 in steps of 0.08. Simple tests using the ML procedure indicate small changes to the step size of our grid are unimportant.

3.2. *HB Template Fitting*

Our HB template fitting method utilizes the same set of SED templates described above in Section 3.1. However, to speed up optimization of the HB classification, we subsampled the ML set of templates to a list of 250 which span a range of physical and color-color properties.

A key set of properties we must specify for our HB approach is the functional form of the prior distributions in the model. Since our templates are discrete both in SED shape and physical properties, we choose the prior distribution to be a weight assigned to each template, with range 0 to 1, such that the weights sum to 1. These weights themselves become parameters in our optimization, thus we have 281 parameters corresponding to template priors since we use 250 star and 31 galaxy templates. We also allow the overall prior distribution that any given object is S or G to be parameterized as two weights, which we optimize.

Finally, we must decide a form for our redshift priors. Ideally, these two should be weights for each discrete redshift, repeated as a separate set for each galaxy template. Unfortunately, this would not only add 51×31 more parameters to optimize, but also significantly slows down likelihood computations. Instead, we adopt a flat prior distribution across redshifts. While not ideal, such a prior eases comparison with ML classification results, and eliminates the need to specify an informative prior which correctly describes the data. Tests of flat versus fixed-form prior distributions indicate that the classification results presented in Section 5 do not vary substantially between the two choices. In total, we optimize the 283 (hyper)parameters of our priors to the maximum likelihood of the entire dataset.

3.3. *SVM Models*

We use the `LIBSVM`[6] set of routines, described in **?**. The provided routines are quick and easy to implement, but require the user to specify a training set of data, as well as the form and parameter values of the kernel function used.

To select the training data, we consider two scenarios. First is a 'best case' situation (SVM$_{best}$), where a random sample of data are selected as a training set. Second, we consider a case where only a high signal–to–noise (S/N) sample of data is available for training (SVM$_{high\,S/N}$). We consider SVM$_{best}$ a optimistic scenario – obtaining a large spectroscopic or multi-wavelength sample of training data, down to the limiting magnitude of a given survey, is very costly in terms of telescope time. The other extreme, SVM$_{high\,S/N}$, is a bit more realistic – for a given survey, obtaining classifications are typically only easily obtained at the high S/N end of the data. In both cases, we consider a training sample size which is a fifth of the total data.

We implement a Gaussian radial basis function for the SVM kernel, with form $k(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\gamma|\mathbf{x_i} - \mathbf{x_j}|^2)$ and an error penalty of $C_{\rm SVM}$. To find the appropriate values of the nuisance parameters $\gamma, C_{\rm SVM}$, we initially explore the classification performance on a grid of $\gamma, C_{\rm SVM}$ values. Ultimately, we optimize to find the best performing values.

[4] We use apparent r magnitude here.

[5] `http://www.cfht.hawaii.edu/%7Earnouts/LEPHARE/lephare.html`

[6] `http://www.csie.ntu.edu.tw/%7Ecjlin/libsvm/`

4. TEST DATA

In order to demonstrate the advantages and disadvantages of star–galaxy classification techniques, we need a test dataset which has a large number of sources, a well understood and calibrated catalog, and for which spectroscopy or multi-wavelength observations reveal the true source classification. In addition, we want these data to be magnitude limited as faint as $r \geq 24$ in order to understand the problem of classification in current and upcoming surveys like Pan-STARSS, DES, and LSST.

To satisfy these requirements, we utilize the COSMOS dataset as a test case for photometric classification. The COSMOS survey (Scoville et al. 2007) covers ~ 2 square degrees on the sky using 30 band photometry, and is magnitude limited down to $r \sim 25$. Broadband $ugrizJK$ photometry exists down to limiting magnitudes which complement the r limiting magnitude, and *Spitzer* IRAC coverage exist for sources as faint as $K \lesssim 24$. In addition, *GALEX* and *XMM* coverage are of sufficient depth to pick out relatively bright star–forming galaxies and AGN. The spectral coverage beyond the optical, particularly the near–infrared, can be a powerful discriminator between star and galaxy classification. For instance, Ilbert et al. (2009) show the $r - m_{3.6\mu\,\mathrm{m}}$ vs. $r - i$ colors cleanly separate star and galaxy loci, since stars have systematically lower $r - m_{3.6\mu\,\mathrm{m}}$ colors. In addition to 30 band photometry, the COSMOS field has deep *HST* $i-$band coverage, down to a limiting magnitude of $i \sim 28$. Diffraction limited *HST* imaging allows the morphological discrimination of point-like and extended sources, further strengthening the fidelity of the COSMOS star–galaxy classification.

We follow the COSMOS team's star–galaxy classification criteria in order to determine the 'true' classification for the purpose of testing our methods. These consist of a χ^2 classification from fitting star and galaxy templates to the 30 band photometry, and a morphological classification using the `ACS_MU_CLASS` statistic derived by the analysis of the *HST* photometry by Scarlata et al. (2007). We use an updated version of the publicly available photometric redshift catalog, provided by P. Capak (private communication).

For our final catalog of test data, we choose to restrict our analysis to sources likely to be unresolved in ground based data. We do so since tried and true morphological classification criteria will easily distinguish quite extended sources, accounting for the vast majority of galaxies to depths of $r \sim 24 - 25$. Using the full-width half-maximum (FWHM) angular sizes found with *HST* imaging in COSMOS, we examine sources with FWHM $< 0.2''$. Such sizes are likely never to resolved in surveys with seeing $\gtrsim 0.7''$, even with sophisticated morphological analysis, and are thus an excellent test bed for the type of sources which will rely the most on photometric criteria. In total, our sample consists of 17616 sources with apparent magnitudes $22.5 < r < 25$, and is plotted in *ugriz* color–color space in Figure 2.

5. RESULTS AND DISCUSSION

We report the classification performance of Maximum Likelihood (ML) and Hierarchical Bayesian (HB) template fitting, as well as a thoroughly tested Support Vector Machine (SVM) on our COSMOS based test data.

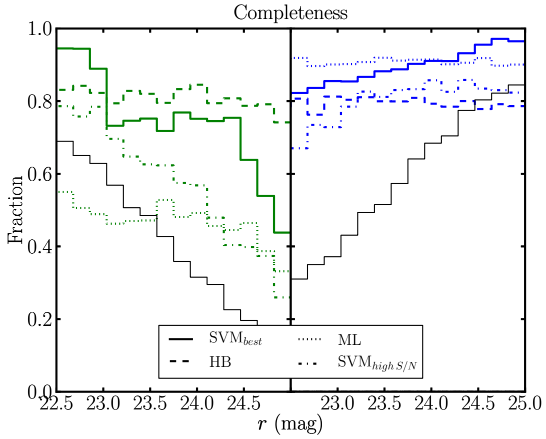

FIG. 3.— The completeness as a function of magnitude produced by the indicated classification approaches. Results for stars are on the left in green, while those for galaxies are shown on the right in blue. The thin, solid black line indicates the sample fraction for a given object type. For galaxies, the various methods return similar completeness values, while the discrepancy is much larger in the case of stars.

There are many different measures which can be used to assess the performance of each algorithm. First, we consider the completeness[7] and purity[8] of classified samples, evaluated at $\ln(\Omega) = 0$. Figures 3 and 4, display the completeness and purity, respectively, as a function of magnitude. Examining Figure 3, it is not immediately clear which classification approach delivers the best completeness. For galaxies, all methods seem to be competitive, returning $80 - 90\%$ completeness across all magnitudes. In the case of stars, however, it is clear only our HB template fitting and SVM$_{best}$ deliver acceptable completeness – at $r > 24$ the completeness of ML template fitting and SVM$_{high\,S/N}$ fall to 50% or below. In terms of purity (Figure 4), SVM$_{best}$ clearly outperforms all other approaches. Indeed, for galaxies, HB and SVM$_{high\,S/N}$ yield similar performance to SVM$_{best}$, but all approaches underperform SVM$_{best}$ in terms of stellar purity. It should be noted that HB delivers similar to or better than performance than ML in all cases, even with the relatively simple HB approach presented here.

It is interesting to note the regions of *ugirz* color–color space where classification fails. In Figures 5 and 6, we show the fraction of sources correctly classified using HB and SVM$_{best}$, distributed over colors. Comparing with Figure 2, it is unsurprising that the places where classification is least successful are regions that overlap the most in color. For example, in the region of $r - i > 1.5$ the stellar locus has essentially zero overlap with galaxies in the sample, and has successful classification for both SVM$_{best}$ and HB approaches. For our HB technique, this is not to say there are no galaxy template models in the region of $r - i > 1.5$ but rather the algorithm infers that there are no galaxies in that region of data. One obvious difference between the two figures is the relative success in $u - g$ color space – HB clearly underperforms SVM$_{best}$. This difference is likely do to two reasons. First, SVM$_{best}$

[7] Defined as the fraction of sources of true type X, correctly classified as X.

[8] Defined as the number of sources of true type X, correctly classified as X, divided by the total number of sources classified as X.

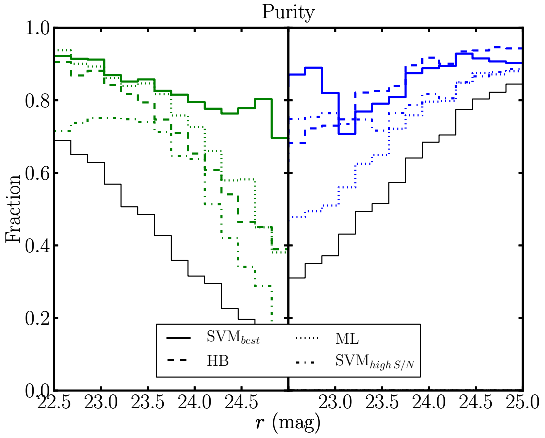

FIG. 4.— Similar to Figure 3 but showing purity of classified samples, instead of completeness. Here, SVM algorithms clearly outperform all others, if given a very good set of training data (SVM$_{best}$). For galaxies, our HB algorithm delivers similar purity to the SVM$_{best}$ scenario. For stars, however, HB underperforms SVM$_{best}$ as the stellar fraction of the sample decreases

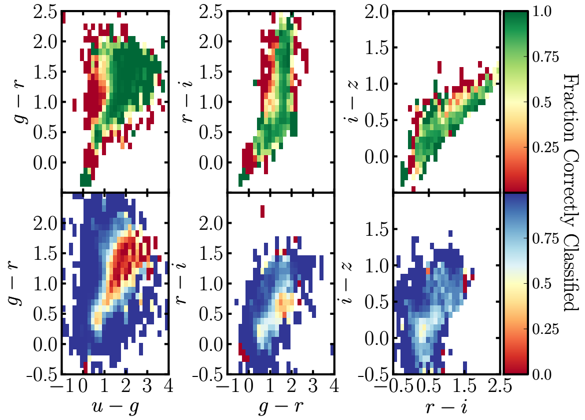

FIG. 5.— The fraction of objects correctly classified at $\ln(\Omega) = 0$ using our HB template fitting, distributed in *ugriz* color-color space. Comparing with Figure 2, it is clear that classification is most successful for regions in which stars and galaxies do not overlap in color-color space.

is data driven, and as such will always have a better model than a sparse set of templates. Second, the number density of galaxies in the failing region is low, making HB more likely to call everything a star. While less obvious in other regions of *ugriz* color space, these reasons are likely extensible to understanding the relatively lower performance of HB versus SVM$_{best}$.

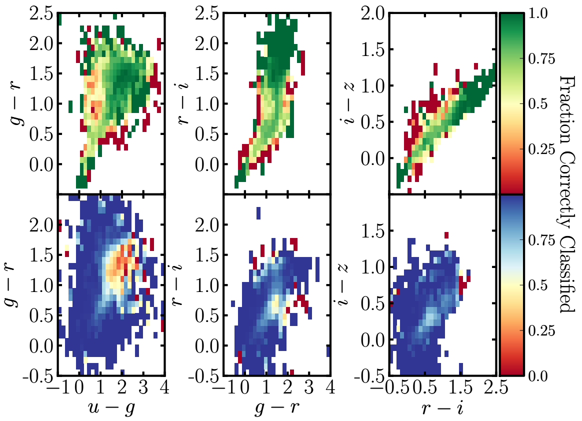

FIG. 6.— The same as Figure 5, but for a SVM trained with data which span the S/N range of the whole sample (SVM$_{best}$). By inspection, it is clear that SVM$_{best}$ outperforms HB template fitting, particularly in the case of galaxies. A striking difference is the poor galaxy classification of HB compared to SVM$_{best}$ in $u-g$. This may indicate a model deficiency in the u spectral range of our galaxy templates.

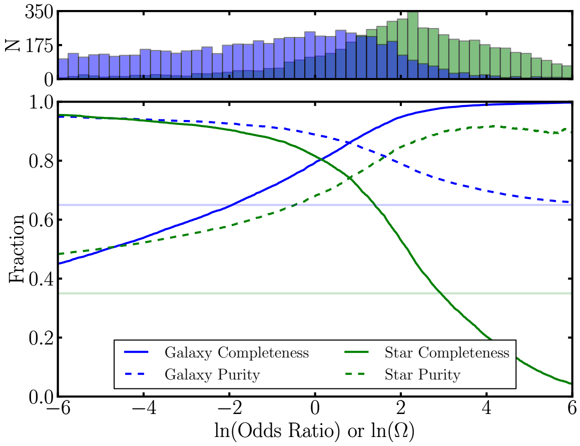

FIG. 7.— Hierarchical Bayesian template fitting results showing completeness (solid) and purity (dashed) lines as a function of $\ln\Omega$. Results for stars are shown in green and galaxies are shown in blue, while the solid (dashed) curves show completeness (purity). Also indicated by green and blue horizontal lines is the relative fraction of stars and galaxies in the sample, respectively. The top panel shows the histograms associated with the distribution. Setting $\ln\Omega >= 6$ effectively calls all sources galaxies, so the galaxy completeness is high, while the purity is set by the sample fraction of galaxies. The same conclusions are reached for stars at $\ln\Omega < -6$. The exact value of $\ln\Omega$ chosen depends on the completeness and purity requirements dictated by the user's science case.

One of the great advantages of probabilistic classification is that one need not restrict the classification criterion to a fixed value. By moving away from $\ln(\Omega) = 0$, one can obtain more/less pure or complete samples of stars and galaxies, depending on the user's science case. In detail, how the purity or completeness varies as a function of $\ln(\Omega)$ depends on the algorithm used. To illustrate, we show in Figure 7 how purity and completeness vary for the log odds ratio output by our HB algorithm. In the figure, as $\ln(\Omega)$ decreases, we are requiring that the relative likelihood that an object is a galaxy is much higher than that for a star. Similarly, as $\ln(\Omega)$ increases we are requiring objects be more stringently classified as a star. Thus, by moving away from $\ln(\Omega) = 0$ we change the star/galaxy purity and completeness to the point where everything is called a star or galaxy, giving 100% complete samples with a purity set by the sample fraction. One caveat, however, is that modifying the threshold Ω to achieve more pure samples may select objects which lie in particular regions in SED space. To illustrate, we show in Figure 8 the distribution of $\ln(\Omega)$ in color space.

We have considered one means of comparing different source classification approaches, examining the completeness and purity of samples (as a function of $\ln(\Omega)$). However, such comparisons are often muddled by the many ways to view purity and completeness. For exam-

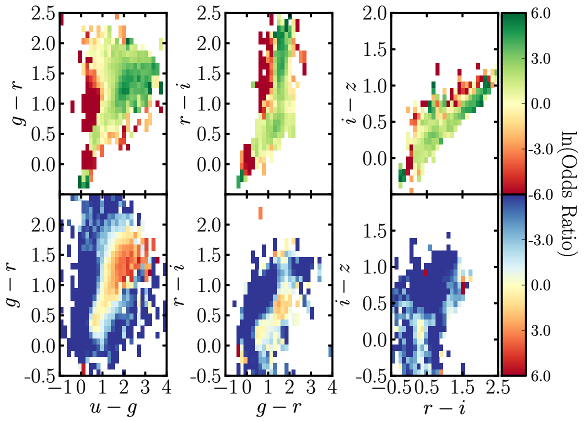

FIG. 8.— The median $\ln(\Omega)$ of objects produced by our HB template fitting, distributed in *ugriz* color-color space. Similar to Figure 5, regions with the most extreme $\ln(\Omega)$ values are primarily those which have little color–color overlap between stars and galaxies. While altering the $\ln(\Omega)$ threshold can deliver more pure or complete samples (cf. Figure 7), it may likely bias the sample to certain regions of color space.

ple, Figure 3 shows that compared to SVM$_{best}$, our HB method gives better completeness in stars but slightly worse completeness for galaxies – which performs better in general?

We assess the overall performance of the various classification approaches using the Receiver Operating Characteristic (ROC) curve. A ROC curve is a plot of the true positive rate versus the false positive rate of a binary classifier, as the classification threshold ($\ln(\Omega)$) is varied. In Figure 9, we plot the ROC curve for all four classification approaches considered here. An ideal classifier has a true positive rate equal to one for all values of $\ln(\Omega)$. Thus, the Area Under the Curve (AUC) statistic is an assessment of the overall performance of the classifier. There are several points worth noting in Figure 9. First, our HB approach to template fitting clearly outperforms the ML approach. Considering our simple HB implementation is not very computationally demanding (tens of minutes on typical desktop computer), even a basic HB approach should always be preferred over the ML case. SVM algorithms, when trained with data which accurately capture the SED and S/N properties of the entire data, generally perform much better than our current template fitting methods. This is not surprising, since template driven algorithms are never likely to have as complete models as something data driven. While this provides encouragment for classification using SVM, a major hurdle for accurate classification is the need for a very good training dataset. Obtaining a training set with accurate classification is difficult, requiring spectroscopy or (at a minimum) multi-wavelength data that spans into the infrared. Since these are expensive to obtain for a large number of sources, especially at depths of $r \gtrsim 24$ and beyond. Instead, it is likely that available training data will only capture the high S/N end of the survey in question. As shown in Figure 9 a SVM$_{high\,S/N}$ scenario underperforms even ML template fitting, casting doubt onto the usefulness of SVM with ill–suited training information.

In addition to uncertainties associated with the adequacy of training data, SVM codes also retain uncertainties associated with kernel nuisance parameters ($\gamma, C_{\rm SVM}$ here). To obtain the best possible SVM results, we tuned these parameters using the known classifications of all our test data. Such a scenario will not be possible for future surveys, and will rely on tuning within the training data, or on experience from much higher S/N cases. In our experience we have found good SVM classification for a wide range of nuisance parameters, but there is no guarantee this should be the case. Our HB template fitting, however, is fully specified - all nuisance and hyperparameters are determined by the optimum likelihood for all the data.

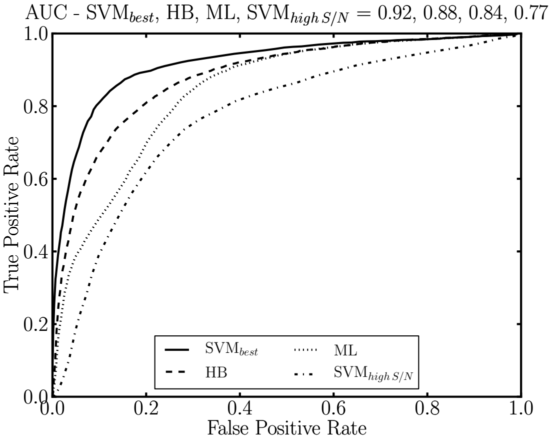

FIG. 9.— The Receiver Operating Characteristic (ROC) curve for four photometric classification approaches: SVM$_{best}$, SVM$_{high\,S/N}$, ML, and HB. The ROC curve shows the true positive rate versus the false positive rate, as $\ln(\Omega)$ varies. An ideal classifier always returns a true positive rate of one, so the Area Under the Curve (AUC) provides a general assessment of the performance.

6. CONCLUSIONS

Imminent and upcoming ground–based surveys are observing large portions of the sky in optical filters to depths ($r \gtrsim 24$), investing large amounts of money, resources, and (wo)manpower. In order for such surveys to best achieve many of their science goals, accurate star–galaxy classification is required. At these new depths, unresolved galaxy counts increasing dominate the number of point sources classified through morphological means. To investigate the usefulness of photometric classification methods for unresolved sources, we examine the performance of photometric classifiers using *ugriz* photometry of COSMOS sources with intrinsic FWHM $< 0.2''$, as measured with *HST*. Our conclusions are as follows:

- Maximum Likelihood (ML) template fitting methods are simple, and return informative classifications. At $\ln(\Omega) = 0$, ML methods deliver high galaxy completeness ($\sim 90\%$) but low stellar completeness ($\sim 50\%$). The purity of these samples range from $\sim 50-90\%$, and are a strong function of the relative sample fraction.

- We present a new Hierarchical Bayesian (HB) approach to template fitting which outperforms ML techniques, as shown by the Receiver Operating Characteristic (ROC). Our HB algorithm currently optimizes the likelihood by only modifying prior weights for template models. Further improvements are possible by hierarchically modeling the redshift distribution of galaxies, the SEDs of the input templates, and the distribution of apparent magnitudes.

- Support Vector Machine (SVM) algorithms can deliver excellent classification, which outperforms template fitting methods. Successful SVM performance relies on having an adequate set of training data. For optimistic cases, where the training data is essentially a random sample of the data (with known classifications), SVM will outperform template fitting. In a pessimistic scenario, where the training data is from higher signal to noise than the whole sample, SVM algorithms perform worse than the simplest template fitting methods.

- HB algorithms have no need for training, and have nuisance parameters that are tuned according to the likelihood of the data itself. Since it is unclear when, if ever, adequate training data will be available for classification, HB algorithms may prove a to be a useful approach for next-generation classifiers.

We gratefully acknowledge P. Capak and the COSMOS team for providing an up–to–date version of their catalog, C.-C. Chang and C.-C. Lin for making their SVM routines available, and the `Le Phare` photo-z team for making code and templates available. We also thank J. Newman, [**RF: insert others??**] for helpful and insightful conversations. [**RF: insert grant info**]

REFERENCES

????

08. 1

Arnouts, S., Cristiani, S., Moscardini, L., Matarrese, S., Lucchin, F., Fontana, A., & Giallongo, E. 1999, MNRAS, 310, 540

Bochanski, J. J., West, A. A., Hawley, S. L., & Covey, K. R. 2007, AJ, 133, 531

Castelli, F., & Kurucz, R. L. 2004, ArXiv Astrophysics e-prints

Chang, C.-C., & Lin, C.-J. 2011, ACM Transactions on Intelligent Systems and Technology, 2, 27:1, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. 2003, Bayesian Data Analysis, Second Edition (Chapman & Hall/CRC Texts in Statistical Science), 2nd edn. (Chapman & Hall)

Hogg, D. W., Bovy, J., & Lang, D. 2010, arXiv:1008.4686

Ilbert, O. et al. 2006, A&A, 457, 841

Ilbert, O., et al. 2009, ApJ, 690, 1236

Le Fèvre, O. et al. 2005, A&A, 439, 877

Pickles, A. J. 1998, PASP, 110, 863

Polletta, M. et al. 2007, ApJ, 663, 81

Saglia, R. P. et al. 2012, ApJ, 746, 128

Scarlata, C. et al. 2007, ApJS, 172, 406

Scoville, N. et al. 2007, ApJS, 172, 1

Shu, Y., Bolton, A. S., Schlegel, D. J., Dawson, K. S., Wake, D. A., Brownstein, J. R., Brinkmann, J., & Weaver, B. A. 2012, AJ, 143, 90

APPENDIX

HIERARCHICAL BAYESIAN STAR–GALAXY CLASSIFICATION

Let us define the data as the sets:

$$\begin{aligned}\boldsymbol{F} &= \{10^{-\frac{2}{5}m_1}F_{1,0}, \dots, 10^{-\frac{2}{5}m_l}F_{l,0}, \dots, 10^{-\frac{2}{5}m_N}F_{N,0}\}\\ \boldsymbol{\sigma_F} &= \{\frac{2}{5}\ln(10)F_1\sigma_{m_1}, \dots, \frac{2}{5}\ln(10)F_l\sigma_{m_l}, \dots, \frac{2}{5}\ln(10)F_N\sigma_{m_N}\} \quad ,\end{aligned} \tag{A1}$$

where m_l, σ_{m_l} is the observed magnitude and uncertainty in filter number l for N number of filters. One sequence for the filters l correspond to $\{l\} = \{u, g, r, i, z\}$. The zeropoint, $F_{l,0}$, is:

$$F_{l,0} = \int \lambda S_\lambda R_{\lambda,l} \mathrm{d}\lambda \quad , \tag{A2}$$

where S_λ is the standard flux density spectrum (Vega or AB), and $R_{\lambda,i}$ is the fraction of photons incident on the top of the atmosphere which are counted by the detector, as a function of wavelength.

Next, we generate a model for the data using the templates:

$$F_{\mathrm{mod},l} = \int \lambda f_{\lambda,\mathrm{mod}} R_{\lambda,l} \mathrm{d}\lambda \quad , \tag{A3}$$

where $f_{\lambda,\mathrm{mod}}$ corresponds to the flux density of a given spectral template. Finally, we define a goodness of fit statistic:

$$\chi^2 = \sum_{l=1}^{N} \frac{(F_l - C_{\mathrm{mod}}F_{\mathrm{mod},l})^2}{\sigma^2_{\mathrm{total}_l}} \quad , \tag{A4}$$

where C_{mod} is a constant unitless amplitude applied to the model for the fit (discussed more below as C_{ij}). The variance $\sigma^2_{\mathrm{total}_l} = \sigma^2_{F_l} + \eta F_l$, where η is a few percent and represents a nuisance parameter which accounts for error in the models as well as underestimates in $\sigma^2_{F_l}$. The value of χ^2 from our template fitting is the fundamental quantity on which our inference procedure is based, as follows below.

We represent the hypothesis that an object i is a star or a galaxy by "S" or "G" respectively. For a given object i, we fit a set of templates j corresponding to S using the procedure outlined above. The likelihood of template j and amplitude (flux, or brightness, or inverse-squared distance) C_{ij} under the stellar hypothesis S given the single observed data point $\boldsymbol{F}_i$ is:

$$p(\boldsymbol{F}_i|C_{ij}, j, S) \propto \exp(-\frac{1}{2}\chi^2) \quad , \tag{A5}$$

where $\boldsymbol{F}_i$ is the full set of observations of object i and the associated noise model, and χ^2 is defined in Equation A4. Note that the χ^2 is not necessarily the best-fit value for χ^2 but rather the χ^2 obtained with template j when it is given amplitude C_{ij}.

We could optimize this likelihood, but really we want to compare the whole S model space to the whole G model space. We must marginalize this likelihood over the amplitude and template. To demonstrate this, let us step through each marginalization for the S model space.

Marginalization over the amplitude C_{ij} looks like

$$p(\boldsymbol{F}_i|j, S, \boldsymbol{\alpha}) = \int p(\boldsymbol{F}_i|C_{ij}, j, S)\, p(C_{ij}|j, S, \boldsymbol{\alpha})\, \mathrm{d}C_{ij} \quad , \tag{A6}$$

where the integral is over all permitted values for the amplitude C_{ij}, and the prior PDF $p(C_{ij}|j, S, \boldsymbol{\alpha})$ depends on the template j, the full hypothesis S. Note, the prior PDF obeys the normalization constraint

$$1 = \int p(C_{ij}|j, S, \boldsymbol{\alpha})\, \mathrm{d}C_{ij} \quad . \tag{A7}$$

Here we have also introduced some “hyperparameters” $\boldsymbol{\alpha}$, which are variables which parameterize prior distributions. The subset of hyperparameters $\boldsymbol{\alpha}$ which apply to $p(C_{ij}|j,S,\boldsymbol{\alpha})$ might be, e.g., the mean and variance of a log–normal distribution on C_{ij}. It is the simultaneous inference of the star–galaxy probabilities and the hyperparameters that make the approach hierarchical.

Any realistic prior PDF for the C_{ij} comes from noting that (for stars), the C_{ij} are dimensionless squared distance ratios between the observed star and the template star; in this case the prior involves parameters of the stellar distribution in the Galaxy. When we look at galaxies (below), this situation will be different. In the (rare) case that the prior PDF $p(C_{ij}|j,S,\boldsymbol{\alpha})$ varies slowly around the best-fit amplitude,

$$p(\boldsymbol{F}_i|j,S,\boldsymbol{\alpha}) \propto \exp(-\frac{1}{2}\,\tilde{\chi}^2)\,p(\tilde{C}_{ij}|j,S,\boldsymbol{\alpha})\,\sigma_{Cij} \quad , \tag{A8}$$

where $\tilde{\chi}^2$ is the best-fit chi-squared, $\tilde{C}_{ij}$ is the best-fit amplitude, and σ_{Cij} is the standard uncertainty in $\tilde{C}_{ij}$ found by least-square fitting. This approximation is that the prior doesn't vary significantly within a neighborhood σ_{Cij} of the best-fit amplitude.

Marginalization over the template space looks like

$$p(\boldsymbol{F}_i|S,\boldsymbol{\alpha}) = \sum_j p(\boldsymbol{F}_i|j,S)\,P(j|S,\boldsymbol{\alpha}) \quad , \tag{A9}$$

where $P(j|S,\boldsymbol{\alpha})$ is the prior probability (a discrete probability, not a PDF) of template j given the hypothesis S and the hyperparameters $\boldsymbol{\alpha}$. It obeys the normalization constraint

$$1 = \sum_j P(j|S,\boldsymbol{\alpha}) \quad . \tag{A10}$$

Note $P(j|S,\boldsymbol{\alpha})$ is a discrete set of weights, whose value corresponds to the hyperparameter for template j.

To summarize, the marginalized likelihood $p(\boldsymbol{F}_i|S,\boldsymbol{\alpha})$ that a source i is a star S is computed as:

$$\begin{aligned} p(\boldsymbol{F}_i|C_{ij},j,S) &\propto \exp(-\frac{1}{2}\,\chi^2) \\ p(\boldsymbol{F}_i|j,S,\boldsymbol{\alpha}) &= \int p(\boldsymbol{F}_i|C_{ij},j,S)\,p(C_{ij}|j,S,\boldsymbol{\alpha})\,\mathrm{d}C_{ij} \\ p(\boldsymbol{F}_i|S,\boldsymbol{\alpha}) &= \sum_j p(\boldsymbol{F}_i|j,S,\boldsymbol{\alpha})\,P(j|S,\boldsymbol{\alpha}) \quad . \end{aligned} \tag{A11}$$

The marginalized likelihood that source i is a galaxy G, is calculated following a very similar sequence. In calculating the likelihood, we allow a given galaxy template k to be shifted in wavelength by a factor $1+z$. This introduces another step in the calculation that marginalizes the likelihood across redshift for a template, giving

$$\begin{aligned} p(\boldsymbol{F}_i|C_{ikz},k,z,G) &\propto \exp(-\frac{1}{2}\,\chi^2) \\ p(\boldsymbol{F}_i|k,z,G,\boldsymbol{\alpha}) &= \int p(\boldsymbol{F}_i|C_{ikz},k,z,G)\,p(C_{ikz}|k,z,G,\boldsymbol{\alpha})\,\mathrm{d}C_{ikz} \\ p(\boldsymbol{F}_i|G,k,\boldsymbol{\alpha}) &= \sum_z p(\boldsymbol{F}_i|k,z,G)\,P(z|k,G,\boldsymbol{\alpha}) \\ p(\boldsymbol{F}_i|G,\boldsymbol{\alpha}) &= \sum_k p(\boldsymbol{F}_i|k,G)\,P(k|G,\boldsymbol{\alpha}) \quad , \end{aligned} \tag{A12}$$

where now C_{ikz} is the constant amplitude for galaxy template k at a redshift z. The marginalization across redshift also introduces a prior $P(z|k,G,\boldsymbol{\alpha})$, which is also is parameterized by a subset of $\boldsymbol{\alpha}$, under some assumed form for the prior.

This model is fully generative; it specifies for any observed flux $\boldsymbol{F}_i$ the PDF for that observation given either the star hypothesis S or the galaxy hypothesis G. We can write down then the full probability for the entire data set of all objects i:

$$p(\{\boldsymbol{F}_i\}\,|\boldsymbol{\alpha}) = \prod_i \left[p(\boldsymbol{F}_i|S,\boldsymbol{\alpha})\,p(S|\boldsymbol{\alpha}) + p(\boldsymbol{F}_i|G,\boldsymbol{\alpha})\,p(G|\boldsymbol{\alpha})\right] \quad , \tag{A13}$$

where even the overall prior probability $p(S|\boldsymbol{\alpha})$ that an object is a star (or, conversely, a galaxy) depends on the hyperparameters $\boldsymbol{\alpha}$. These obey the normalization constraint

$$1 = p(S|\boldsymbol{\alpha}) + p(G|\boldsymbol{\alpha}) \quad . \tag{A14}$$

The likelihood $p(\{\boldsymbol{F}_i\}\,|\boldsymbol{\alpha})$ is the total, marginalized likelihood for the combined data set of all the observations $\boldsymbol{F}_i$ for all objects i. From here we can take a number of approaches. One option is to find the hyperparameters that maximize this total marginalized likelihood, or we can assign a prior PDF $p(\boldsymbol{\alpha})$ on the hyperparameters, and sample the posterior PDF in the hyperparameter space. For computational reasons, we choose to optimize $p(\{\boldsymbol{F}_i\}\,|\boldsymbol{\alpha})$ in this work.

With either a maximum-likelihood set of hyperparameters $\boldsymbol{\alpha}$ or else a sampling, inferences can be made. For our purposes, the most interesting inference is, for each object i, the posterior probability ratio (or odds) Ω_i

$$\begin{aligned}\Omega_i &\equiv \frac{p(S|\boldsymbol{F}_i,\boldsymbol{\alpha})}{p(G|\boldsymbol{F}_i,\boldsymbol{\alpha})}\\ p(S|\boldsymbol{F}_i,\boldsymbol{\alpha}) &= p(\boldsymbol{F}_i|S,\boldsymbol{\alpha})\,p(S|\boldsymbol{\alpha})\\ p(G|\boldsymbol{F}_i,\boldsymbol{\alpha}) &= p(\boldsymbol{F}_i|G,\boldsymbol{\alpha})\,p(G|\boldsymbol{\alpha}) \quad ,\end{aligned} \tag{A15}$$

where we have re-used most of the likelihood machinery generated (above) for the purposes of inferring the hyperparameters. That is, the star–galaxy inference and the hyperparameter inferences proceed simultaneously.