

## STAR-GALAXY CLASSIFICATION IN MULTI-BAND OPTICAL IMAGING

ROSS FADELY<sup>1</sup>, DAVID W. HOGG<sup>2,3</sup>, & BETH WILLMAN<sup>1</sup>

*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 (intrinsic 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<sub>best</sub>), where the training data is a random sampling of the data in question. In addition, we consider a more realistic scenario where the SVM is trained only using the highest signal-to-noise (S/N) data (SVM<sub>real</sub>). 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 use 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, 0.70 for SVM<sub>best</sub>, HB, ML, and SVM<sub>real</sub>. 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.

*Subject headings:*

### 1. INTRODUCTION

Until now, the primary way that stars and galaxies have been classified in large sky surveys has been a morphological separation (e.g., Kron 1980; Yee 1991; Vasconcellos et al. 2011; Henrion et al. 2011) of point sources (presumably stars) from resolved sources (presumably galaxies). At bright apparent magnitudes, relatively few galaxies will contaminate a point source catalog and relatively few stars will contaminate a resolved source catalog, making morphology a sufficient metric for classification. However, resolved stellar science in the current and next generation of wide-field, ground-based surveys is being challenged by the vast number of unresolved galaxies at faint apparent magnitudes.

To demonstrate this challenge for studies of the field Milky Way (MW), we compare the number of stars to the number of unresolved galaxies at faint apparent magnitudes. Figure 1 shows the fraction of COSMOS sources that are stars as a function of  $r$  magnitude. The COSMOS catalog  $((l, b) \sim (237, 43)$  degrees Capak et al. 2007a; Scoville et al. 2007a; Ilbert et al. 2009) relies on 30-band photometry plus HST/ACS morphology for source classification (see Section 4 for details). Each colored line

shows a different minimum intrinsic full-width half max (FWHM, as measured with HST/ACS imaging) that a source must have to be resolved.

We separately consider relatively bluer ( $g - r < 1.0$ ) and redder ( $g - r > 1.0$ ) sources. Bluer stars are representative of the old, metal-poor main sequence turnoff (MSTO) stars generally used to trace the MW's halo while redder stars are representative of the intrinsically fainter red dwarf stars generally used to trace the MW's disk. We will see that the effect of unresolved galaxies on these two populations is different, both because of galaxy demographics and because the number density of halo MSTO stars decreases at faint magnitudes while the number density of disk red dwarf stars increases at faint magnitudes.

In an optimistic scenario where galaxies with FWHM  $\gtrsim 0.2''$  can be morphologically resolved (the light blue line in Figure 1, third from the top), unresolved galaxies will still greatly outnumber field MW stars in a point source catalog. For studies of blue stars, field star counts are dominated by unresolved galaxies by  $r \sim 22.5$  and are devastated by unresolved galaxies at fainter magnitudes. The problem is far less severe for studies of red stars, which may dominate point source counts for  $r \lesssim 24$ . Although morphological identification of galaxies with FWHM as small as  $0.2''$  is better than possible the Sloan Digital Sky Survey, which had a median seeing of  $\sim 1.3''$ , future surveys with higher median image quality (e.g.  $0.7''$  predicted for LSST) may approach this limit.

Utilizing the fundamental differences between SEDs of

<sup>1</sup> Haverford College, Department of Physics and Astronomy, 370 Lancaster Ave., Haverford, PA, 19041

<sup>2</sup> Center for Cosmology and Particle Physics, Department of Physics, New York University, 4 Washington Place, New York, NY 10003, USA

<sup>3</sup> Max-Planck-Institut für Astronomie, Königstuhl 17, 69117 Heidelberg, Germany

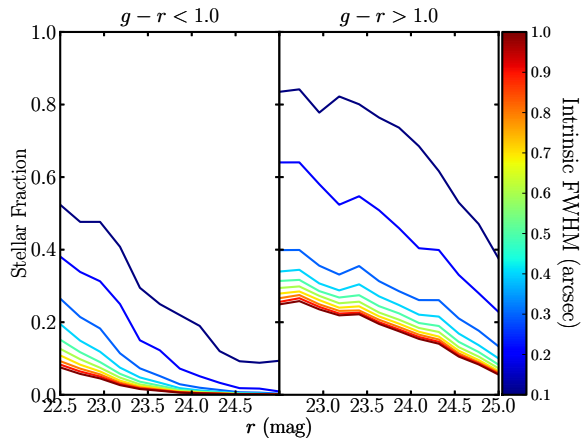


FIG. 1.— The stellar fraction of COSMOS sources as a function of magnitude, for sources with  $g - r < 1$  (left) and  $g - r > 1$  (right). Only stars and galaxies were included in this figure; Only a few percent of the COSMOS point sources are AGN. Colored curves indicate the upper limit in intrinsic full-width half-maximum (FWHM) allowed in the sample. Even in an optimistic scenario where galaxies with  $\text{FWHM} \gtrsim 0.2''$  can be morphologically distinguished from stars, unresolved galaxies will far outnumber stars in point source catalogs at faint magnitudes. This challenge is much greater for blue stars than for red stars.

stars and galaxies can mitigate the contamination of unresolved galaxies in point source catalogs. In general, stellar SEDs are more sharply peaked (close to black-body) than galaxies, which exhibit fluxes more broadly distributed across wavelength. Traditionally, color-color cuts have been used to eliminate galaxies from point source catalogs (e.g. Gould et al. 1992; Reitzel et al. 1998; Daddi et al. 2004). Advantages of the color-color approach include its simple implementation and its flexibility to be tailored to the goals of individual studies. Disadvantages of this approach include its disregard for A and B.

Probabilistic algorithms offer a more general and informative approach to photometric classification. The goal of probabilistic photometric classification of an astronomical source is to use its observed fluxes  $\mathbf{F}$  to compute the likelihood that the object is of a given type. For example, a star ( $S$ ) galaxy ( $G$ ) classification algorithm produces the likelihoods  $p(\mathbf{F}|S)$  and  $p(\mathbf{F}|G)$  and decides classification by comparing the ratio of the likelihoods -

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

A natural classification threshold is an odds ratio,  $\Omega$ , of 1, which may be modified to obtain more pure or more complete samples.

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. Bayesian and  $\chi^2$  template fitting methods have been extensively used to infer the properties of galaxies (e.g. Coil et al. 2004; Ilbert et al. 2009; Xia et al. 2009; Walcher et al. 2011; Hildebrandt et al. 2010). But relatively little attention has been paid to stars which contribute marginally to overall

source counts (although see Robin et al. 2007). Several groups have recently investigated data driven, support vector machine based star-galaxy separation algorithms (e.g. Saglia et al. 2012; Solarz et al. 2012; Tsalmantza et al. 2012).

In this paper, we describe, test, and compare two physically based template fitting approaches to star-galaxy separation (maximum-likelihood and hierarchical bayesian), and one data driven (support vector machine) approach. In Section 2, we present the conceptual basis for each of the three methods. In Section 4, we discuss the specific details, choices, and assumptions made for each of our classification methods. In Section 4, we describe the COSMOS data set with which we test the algorithms. Finally, in Section 5 we show the performance of the algorithms, and discuss the advantages and limitations related to their use as classifiers.

## 2. PROBABILISTIC PHOTOMETRIC CLASSIFICATION TECHNIQUES

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

One common method for inferring an observed source's properties from observed fluxes is template fitting. This method requires a set of spectral templates (empirical or theoretical) that span the possible spectral energy distributions (SEDs) of observed sources. These template SEDs must each cover the full wavelength range spanned by the photometric filters used to measure the fluxes to be fit. The relative template flux in each filter (for example  $ugriz$ ) for each SED is computed by convolving each SED with each filter response curve. Once these template relative flux values are computed for each SED template, the template model is fully specified except for a normalization constant  $C$ . For a given observed source  $i$ , the 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 but can be 'fit' to the data.

The maximum likelihood (ML) value of  $C_i$  for each template that best fits a source's observed fluxes,  $\mathbf{F}$ , is that which returns the lowest  $\chi^2$ . After assessing the ML values of  $C_i$  for all the templates, classification is straightforward - one need only to compare the lowest star  $\chi^2$  to the lowest galaxy  $\chi^2$ . In other words,  $\chi_S^2 - \chi_G^2 = \ln(\Omega)$  is the classification criteria (cf. Equation 1).

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

Hierarchical Bayesian (HB) algorithms provide another template fitting-based approach to photometric classification. Unlike ML approaches, Bayesian approaches offer the opportunity to utilize information about how likely a source is to be  $S$  or  $G$ . With a hierarchical Bayesian algorithm, these probabilities do not need to be set prior to the classification process. We discuss these basic features of Bayesian, and hierarchical Bayesian, algorithms in detail below.

Consider the scenario where a  $G$  model fits data  $\mathbf{F}_i$  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 the wrong thing to do, since the data are stating that  $S$  models are *generally* more favored. Capturing this information is one primary aim of most Bayesian algorithms.

To capture this information, we *marginalize* over all possible star and galaxy templates to compute the total probability that a source belongs to a certain classification ( $S$  or  $G$ ). For a template fitting-based Bayesian algorithm, this marginalization consists of summing up the likelihood of each  $S$  template given  $\mathbf{F}_i$ , as well as the likelihood of each  $G$  template (across redshift). The likelihood of each template was itself calculated as a marginalized likelihood. For each template fit, we compute the total likelihood of the fit by marginalizing over the uncertainty in fitting coefficient  $C_i$ . This marginalization is the total probability of a Gaussian distribution with variance  $\sigma_{C_i}^2$  - a value which is returned using least squares fitting techniques (e.g., Hogg et al. 2010).

By Bayes' theorem, marginalization requires we specify the prior probability that any object might have a given SED template (at a given redshift). The prior probability 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 statistical data analysis (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 exhibit improved probabilistic performance over simpler approaches [RF: Need cites - Hogg, any suggestions???], and require no additional knowledge outside the observed data and the template SEDs. [RF: @BW how about now?] Functionally, hierarchical approaches consist of parameterizing the prior probability distributions (e.g., with 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, we optimize the hyperparameters of the SED template prior distributions to return the maximum marginalized likelihood of all the data. This procedure will enable us to simultaneously infer the star-galaxy probability of each source while determining the hyperparameters that maximize the likelihood of the observed dataset. A brief description of the functional form of these priors is given below in Section 4.2. Although we focus on the star-galaxy probabilities in this paper, the optimized hyperparameters themselves yield a measurement of the detailed demographics of a dataset.

### 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 a gold standard against which to compare any new classification method. SVM algorithms train on a catalog of objects with known classifications to learn the high dimensional boundary that best separates two or more classes of objects. For the case of star-galaxy separation based on broad band photometry, the SVM algorithm learns the boundary which best separates the observed colors and apparent magnitudes<sup>4</sup> of stars and galaxies. For more details on the SVM technique, please see [RF: where to send people??].

<sup>4</sup> We use apparent  $r$  magnitude here.

Successful implementation of a SVM algorithm requires a training set that is a sufficient analog to the set to be classified. A SVM has recently been applied to source classification in the Pan-STARRS 1 photometric pipeline (Saglia et al. 2012), with promising initial results. However, these results were obtained based on analysis of bright, high signal-to-noise data ( $r \lesssim 18$ ), using training data which is a subset of the data itself. To investigate the impact of training set quality and demographics on the problem of star-galaxy separation, we will 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 3), and the number of unresolved galaxies is comparable or more than the number of stars.

### 3. TEST DATA

To investigate the advantages and disadvantages of star-galaxy classification techniques, we need a test catalog which has a large number of sources, is well understood and calibrated, and for which spectroscopy or multi-wavelength observations reveal the true source classifications. 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-STARRS 1, DES, and LSST. The COSMOS catalog satisfies these requirements.

The COSMOS survey (Scoville et al. 2007b) covers  $\sim 2$  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  $K \lesssim 24$  (Capak et al. 2007b; Sanders et al. 2007; Taniguchi et al. 2007). In addition, *GALEX* and *XMM* coverage are of sufficient depth to pick out relatively bright star-forming galaxies and AGN (Hasinger et al. 2007; Zamojski et al. 2007). 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  $r - m_{3.6\mu\text{m}}$  vs.  $r - i$  colors cleanly separate star and galaxy loci, since stars have systematically lower  $r - m_{3.6\mu\text{m}}$  colors. In addition to 30 band photometry, the COSMOS field has *HST/ACS*  $i$ -band coverage, down to a limiting magnitude of  $i \sim 28$  (Koekemoer et al. 2007; Scoville et al. 2007a). 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  $\chi^2$  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). AGN were classified using a similar criterion, as well as the presence of X-ray emission in the XMM data. We use an updated version of the publicly available photometric redshift catalog, provided by P. Capak (private communication).

Throughout this paper, we restrict our analysis to sources likely to be unresolved in ground based data

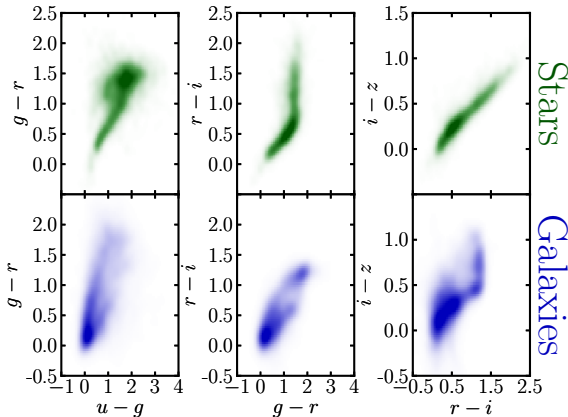


FIG. 2.— The color-color space distribution of point sources ( $\text{FWHM} < 0.2''$ ) in the COSMOS catalog. It is clear that stars in the sample follow a tight locus in all slices of color-color space, while galaxies are more generally distributed. Even so, comparison by eye reveals significant overlap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

( $\text{FWHM}_{HST/ACS} < 0.2''$ ). We do so since tried and true morphological classification criteria will easily distinguish quite extended sources, accounting for a majority of galaxies to depths of  $r \sim 24-25$ . However, galaxies with angular sizes  $< 0.2''$  are unlikely to be resolved in surveys with seeing  $\gtrsim 0.7''$ , so are an appropriate test bed for the type of sources which will rely the most on photometric star-galaxy separation. In total, our sample consists of 6166 stars and 11449 galaxies with apparent magnitudes  $22.5 < r < 25$ , and is plotted in  $ugriz$  color-color space in Figure 2. Over this magnitude range, the median signal-to-noise in the  $r$  band ranges from  $\sim 50$  at  $r = 22.5$  to  $\sim 15$  at  $r = 25$ , with lower corresponding ranges of 10 to 7 in the  $u$ . Of all YY sources with  $\text{FWHM} < 0.2''$  in the COSMOS catalog, we identified XX AGN. Thus, the contamination from any mis-classified AGN should be at the sub-percent level for the sample we use here.

#### 4. IMPLEMENTATION OF THREE STAR-GALAXY CLASSIFIERS

In this Section, we describe our implementation of ML template fitting, HB template fitting, and a SVM on the  $ugriz$  photometry of COSMOS sources for purposes of star-galaxy classification.

##### 4.1. ML Template Fitting

Template based star-galaxy classification relies on the use of spectral energy distribution templates which (as well as possible) span the space of colors for both stars and galaxies. For our stellar model library, we first 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  $-2.5 \leq [\text{Fe}/\text{H}] \leq 0.0$ , surface gravities ranging from  $3.0 \leq \log(g) \leq 0.0$ , and effective temperatures from  $3500 \leq T_{\text{eff}} \leq 10000$  K. We include binary star templates by combining like-metallicity, flux calibrated CK models. Finally, we include 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  $4000 \text{ \AA}$ . We extend these templates down to the  $3000 \text{ \AA}$  using a main sequence CK model with  $T_{\text{eff}} = 3500\text{K}$ . Details of this extension are likely to be unimportant, since the flux of such stars between  $3000 - 4000 \text{ \AA}$  is negligible. Our final combined library of stellar templates includes 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<sup>5</sup> (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.

For all of the above templates, model fluxes were constructed by integrating the SED flux density values with the throughput response curves for each filter. These consist of a  $u^*$  response curve for the observations taken by the Canada-France-Hawaii Telescope, and  $g^+$ ,  $r^+$ ,  $i^+$ ,  $z^+$  response curves for data collected by the Subaru telescope. We obtained the same response curves used by Ilbert et al. (2009) through **Le Phare**<sup>5</sup>. To check for any mismatch between the data, calibrations, and/or response curves, we verified that model colors generated from the SEDs overlap well with the star and galaxy loci.

##### 4.2. HB Template Fitting

While the HB template fitting technique builds on the foundation described in Section 4.1, the details of star-galaxy inference require significantly more mathematical formalism to thoroughly describe. We present the details of this formalism and a detailed, step-by-step description of our HB inferential procedure in Appendix A. Open-source C code is available at <http://github.com/rossfadely/star-galaxy-classification>. In this section, we qualitatively describe features specific to our HB algorithm.

Our HB template fitting method draws from the same set of SED templates described above in Section 4.1. However, to speed up the algorithm, we used only 250 of the 1313 star templates, spanning a range of physical and color-color properties.

The primary choice we must make for our HB approach is the functional form(s) of the prior probability distributions in the model. Since our templates are discrete

<sup>5</sup> <http://www.cfht.hawaii.edu/~Earnouts/LEPHARE/lephare.html>

both in SED shape and physical properties, we parameterize the prior probability of each template to be a single valued weight, within the range 0 to 1, such that the weights sum to 1. These weights themselves become hyperparameters in our optimization. We thus have 281 hyperparameters corresponding to template priors since we use 250 star and 31 galaxy templates. The overall prior probability that any given object is  $S$  or  $G$  is also parameterized as two weights (one for each of  $S$  and  $G$ ), which we optimize.

Finally, we must choose a form for our redshift priors. Ideally, these should be parameterized as weights for each discrete redshift, repeated as a separate set for each galaxy template. Unfortunately, this would not only add  $51 \times 31$  more hyperparameters 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.

#### 4.3. SVM Models

We use the LIBSVM<sup>6</sup>— set of routines, described in Chang & Lin (2011). 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 well-characterized training set exists with both the same object demographics and same S/N as the data to be classified. To emulate this scenario, we select the training set as a random sample of the COSMOS catalog. Second, we consider a more realistic case where the available training set is only sampling the demographics of the high signal-to-noise ( $S/N$ ) portion of the catalog to be classified ( $SVM_{real}$ ). In this case, the demographics of objects in the training set may not match the demographics of the majority of objects in the set to be classified.

We consider  $SVM_{best}$  an 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,  $SVM_{real}$ , is a bit more realistic – for a given survey, classifications are typically easily obtained only 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 catalog size.

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  $C_{SVM}$ . To find the appropriate values of the nuisance parameters  $\gamma, C_{SVM}$ , we initially explore the classification performance on a grid of  $\gamma, C_{SVM}$  values. Ultimately, we optimize to find the best performing values.

Finally, to implement the SVM classification routine we need to scale both the training data and the data to

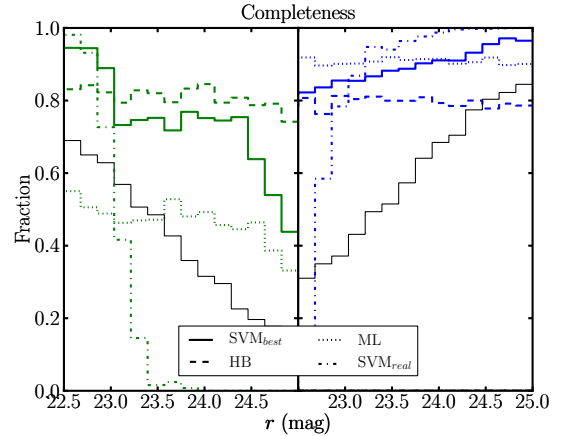


FIG. 3.— The completeness as a function of magnitude produced by the indicated classification approaches. Results for stars are on the left in green, while those for galaxies are shown on the right in blue. The thin, solid black line indicates the sample fraction for a given object type. For galaxies,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

be classified (a.k.a., test data). That is, for the colors and apparent magnitude used, we must scale the range of each to lie between -1 and 1. We map both training and test data to the interval  $[-1, 1]$  using the full range of values in the test data. This is important in the case of  $SVM_{real}$ , since the training data may not span the full range of values for the test data. We find that scaling can have a significant effect for the  $SVM_{real}$  model. For example, poor classification performance is obtained if the  $SVM_{real}$  training data is scaled to itself rather than to the test data

## 5. RESULTS AND DISCUSSION

**[BW: I think we should consider breaking this section up into 2-3 subsections.]**

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. There are many different measures which can be used to assess the performance of each algorithm. First, we consider the completeness<sup>7</sup> and purity<sup>8</sup> of classified samples, evaluated at  $\ln(\Omega)^9 = 0$ . Figures 3 and 4, display the completeness and purity, respectively, as a function of magnitude. Examining Figure 3, all methods seem to be fairly competitive for galaxy classification, returning 80 – 90% completeness across all magnitudes.  $SVM_{best}$  and ML yield the most consistently robust completeness for galaxy. 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 falls to 50% or below, and the completeness for  $SVM_{real}$  goes to zero. The mismatch in source demographics between the realistic training set and the faint COSMOS sources severely undermined the efficacy of the SVM.

<sup>7</sup> Defined as the fraction of sources of true type  $X$ , correctly classified as  $X$ .

<sup>8</sup> Defined as the number of sources of true type  $X$ , correctly classified as  $X$ , divided by the total number of sources classified as  $X$ .

<sup>9</sup> Defined in Equation 1

<sup>6</sup> <http://www.csie.ntu.edu.tw/%7Ecjlin/libsvm/>



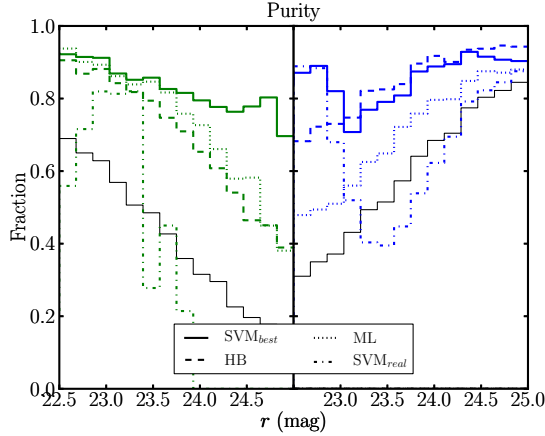


FIG. 4.— Similar to Figure 3 but showing purity of classified samples, instead of completeness. Results for stars are on the left in green, while those for galaxies are shown on the right in blue. Here, SVM algorithms clearly outperform all others, if given a very good set of training data ( $SVM_{best}$ ). 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.

In terms of purity (Figure 4),  $SVM_{best}$  clearly outperforms all other approaches. For galaxies, HB yields similar performance to  $SVM_{best}$ , but all approaches underperform  $SVM_{best}$  in terms of stellar purity. When taking in concert with the results of Figure 3, we see that HB delivers similar to or better than performance than ML in all cases, even with the relatively simple HB approach presented here. For stars, ML and HB yield similar sample purity, but HB does so with a much higher completeness ( $\sim 80\%$  vs.  $\sim 50\%$ ). For galaxies, HB yields a consistently higher sample purity by  $\sim 10 - 15\%$  but a consistently lower sample completeness by  $\sim 15\%$ .

We consider the performance achieved by the  $SVM_{best}$  algorithm to represent the best possible classification of stars and galaxies that could be done, based on *ugriz* photometry alone. However, it is unlikely that an ideal training set will be available for object classification in future, deep datasets. Identifying the regions of *ugriz* color-color space where classification fails can highlight possible ways to improve the HB (or ML) classification methods implemented here, that do not rely on such training sets. For example, we want to check for regions of color-color space in which templates used in ML and HB may be missing, or to check whether the implementation of simple, but stronger, priors could increase performance.

Figures 5 and 6 show the fraction of sources correctly classified using HB and  $SVM_{best}$ , distributed over colors. Comparing with Figure 2 reveals that the places where classification is least successful are regions where stars and galaxies overlap the most in color. For example, both the  $SVM_{best}$  and the HB algorithm struggle to correctly identify galaxies with  $1 < u - g < 3$  and  $1 < g - r < 1.5$ . The number density of galaxies in the failing region is low, making HB even more likely to call everything a star. Similarly, both stars and galaxies populate  $u - g < 1$  and  $g - r \sim 1$ , presenting a challenge to both SVM and HB algorithms. In this case, the number density of galaxies is higher than that of stars, making HB even more likely to call everything a galaxy and training SVM on a color separation that favors galaxies over stars.

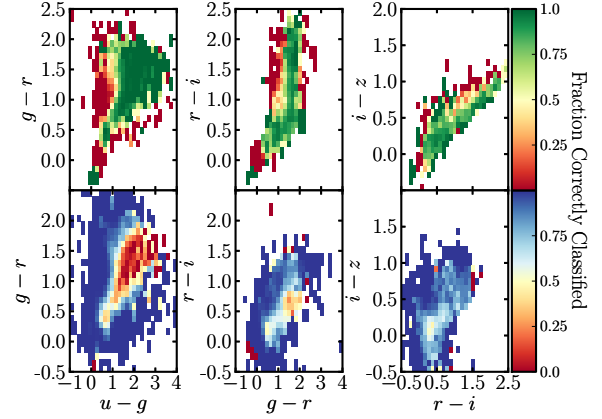


FIG. 5.— The fraction of objects correctly classified at  $\ln(\Omega) = 0$  using our HB template fitting, distributed in *ugriz* color-color space. The top panel shows the performance on stars, and the bottom panel shows the performance on galaxies. Comparing with Figure 2, it is clear that classification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

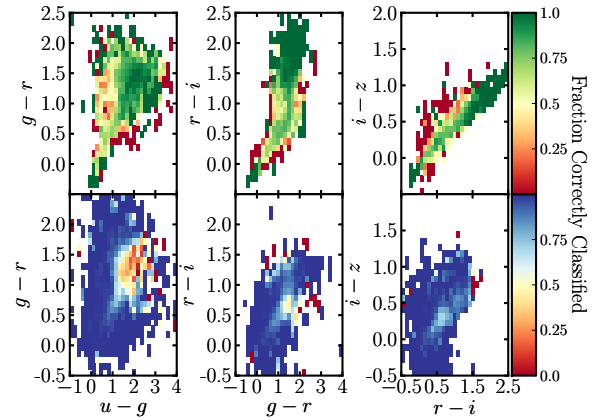


FIG. 6.— The same as Figure 5, but for a SVM trained with data which span the S/N range of the whole sample ( $SVM_{best}$ ). The top panel shows the performance on stars, and the bottom panel shows the performance on galaxies. By inspection, it is clear that  $SVM_{best}$  outperforms HB template fitting, 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.

The region of  $r - i > 1.5$  the stellar locus has essentially zero overlap with galaxies in the sample. The  $SVM_{best}$  algorithm yields exquisite classification of stars in this region of color-color space, which the HB algorithm returns only a mediocre performance (although  $g - r < 1$  and  $r - i > 1.5$  is populated with few stars, so those poorly classified regions do not represent a significant fraction of stars). In future work, the classification of  $r - i > 1.5$  stars could therefore be improved with the implementation of stronger priors on the permitted redshifts at which galaxies may live - for example, by forcing a zero probability of elliptical galaxies at high redshifts.

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 func-

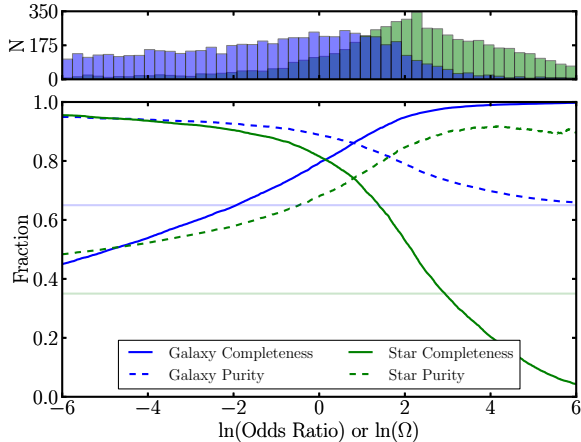


FIG. 7.— Hierarchical Bayesian template fitting results showing completeness (solid) and purity (dashed) lines as a function of  $\ln \Omega$ . Results for stars are shown in green and galaxies are shown in blue, while the solid (dashed) curves show completeness (purity). Also indicated by green and 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  $\ln \Omega \geq 6$  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.

tion 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  $\Omega$  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 the completeness and purity of sets of data classified as stars or galaxies (as a function of  $\ln(\Omega)$ ) as one means of comparing different classification algorithms. A strength of this approach to quantifying the efficacy of classification algorithms is its transparent connection to different science requirements, in terms of purity and completeness. A weakness of this approach is the impossibility of selecting an overall “best” algorithm that presents an average over competing scientific requirements. For example, Figure 3 shows that compared to  $\text{SVM}_{\text{best}}$ , 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 algorithms 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

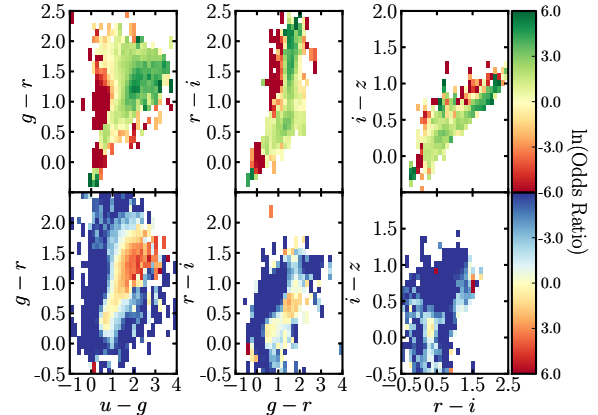


FIG. 8.— The median  $\ln(\Omega)$  of objects produced by our HB template fitting, distributed in  $ugriz$  color-color space. Similar to Figure 5, regions with the most extreme  $\ln(\Omega)$  values are primarily those which have little color-color overlap between stars and galaxies. While altering 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.

$\ln(\Omega)$ . 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 encouragement 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  $\text{SVM}_{\text{real}}$  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 ( $\gamma, C_{\text{SVM}}$  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.

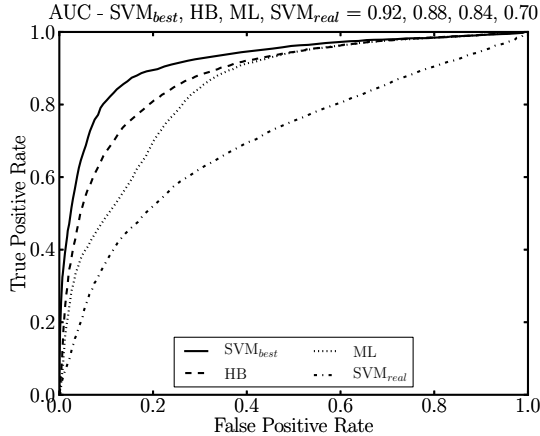


FIG. 9.— The Receiver Operating Characteristic (ROC) curve for four photometric classification approaches:  $SVM_{best}$ ,  $SVM_{real}$ , ML, and HB. The ROC curve shows the true positive rate versus the false positive rate, as  $\ln(\Omega)$  varies. An ideal classifier always returns a true positive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

## 6. CONCLUSIONS

[BW: A point that I want to add to the conclusions is that, although HB and ML do not explicitly require a training set, they DO require a truth table (aka training set) in order to quantify the performance of a classification algorithm. If you don't know your performance, then you suck anyway.]

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 person power. 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 increasingly 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 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, Paul Thorman and his office mate, Mario Juric, [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  
Capak, P. et al. 2007a, ApJS, 172, 99  
—. 2007b, ApJS, 172, 99  
Castelli, F., & Kurucz, R. L. 2004, arXiv:0405087  
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>  
Coil, A. L., Newman, J. A., Kaiser, N., Davis, M., Ma, C.-P., Kocevski, D. D., & Koo, D. C. 2004, ApJ, 617, 765  
Daddi, E., Cimatti, A., Renzini, A., Fontana, A., Mignoli, M., Pozzetti, L., Tozzi, P., & Zamorani, G. 2004, ApJ, 617, 746  
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)  
Gould, A., Guhathakurta, P., Richstone, D., & Flynn, C. 1992, ApJ, 388, 345  
Hasinger, G. et al. 2007, ApJS, 172, 29  
Henrion, M., Mortlock, D. J., Hand, D. J., & Gandy, A. 2011, MNRAS, 412, 2286  
Hildebrandt, H., et al. 2010, arXiv:1008.0658  
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  
Koekemoer, A. M. et al. 2007, ApJS, 172, 196  
Kron, R. G. 1980, ApJS, 43, 305  
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  
Reitzel, D. B., Guhathakurta, P., & Gould, A. 1998, AJ, 116, 707  
Robin, A. C., et al. 2007, ApJS, 172, 545  
Saglia, R. P. et al. 2012, ApJ, 746, 128  
Sanders, D. B. et al. 2007, ApJS, 172, 86  
Scarlata, C. et al. 2007, ApJS, 172, 406  
Scoville, N. et al. 2007a, ApJS, 172, 38  
—. 2007b, 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  
Solarz, A. et al. 2012, ArXiv e-prints  
Taniguchi, Y. et al. 2007, ApJS, 172, 9  
Tsalmantza, P. et al. 2012, A&A, 537, A42  
Vasconcellos, E. C., de Carvalho, R. R., Gal, R. R., LaBarbera, F. L., Capelato, H. V., Frago Campos Velho, H., Trevisan, M., & Ruiz, R. S. R. 2011, AJ, 141, 189  
Walcher, J., Groves, B., Budavári, T., & Dale, D. 2011, Ap&SS, 331, 1  
Xia, L. et al. 2009, AJ, 138, 95  
Yee, H. K. C. 1991, PASP, 103, 396  
Zamojski, M. A. et al. 2007, ApJS, 172, 468

## APPENDIX

## HIERARCHICAL BAYESIAN STAR-GALAXY CLASSIFICATION

Let us define the data as the sets:

$$\mathbf{F} = \{10^{-\frac{2}{5}m_1} F_{1,0}, \dots, 10^{-\frac{2}{5}m_l} F_{l,0}, \dots, 10^{-\frac{2}{5}m_N} F_{N,0}\}$$

$$\sigma_{\mathbf{F}} = \left\{ \frac{2}{5} \ln(10) F_1 \sigma_{m_1}, \dots, \frac{2}{5} \ln(10) F_l \sigma_{m_l}, \dots, \frac{2}{5} \ln(10) F_N \sigma_{m_N} \right\} \quad , \quad (\text{A1})$$

where  $m_l$ ,  $\sigma_{m_l}$  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:

$$F_{l,0} = \int \lambda S_{\lambda} R_{\lambda,l} d\lambda \quad , \quad (\text{A2})$$

where  $S_{\lambda}$  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:

$$F_{\text{mod},l} = \int \lambda f_{\lambda,\text{mod}} R_{\lambda,l} d\lambda \quad , \quad (\text{A3})$$

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

$$\chi^2 = \sum_{l=1}^N \frac{(F_l - C_{\text{mod}} F_{\text{mod},l})^2}{\sigma_{\text{total},l}^2} \quad , \quad (\text{A4})$$

where  $C_{\text{mod}}$  is a constant unitless amplitude applied to the model for the fit (discussed more below as  $C_{ij}$ ). The variance  $\sigma_{\text{total},l}^2 = \sigma_{F_l}^2 + \eta F_l$ , where  $\eta$  is a few percent and represents a nuisance parameter which accounts for error in the models as well as underestimates in  $\sigma_{F_l}^2$ . The value of  $\chi^2$  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  $\mathbf{F}_i$  is:

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

where  $\mathbf{F}_i$  is the full set of observations of object  $i$  and the associated noise model, and  $\chi^2$  is defined in Equation A4. Note that the  $\chi^2$  is not necessarily the best-fit value for  $\chi^2$  but rather the  $\chi^2$  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

$$p(\mathbf{F}_i|j, S, \alpha) = \int p(\mathbf{F}_i|C_{ij}, j, S) p(C_{ij}|j, S, \alpha) dC_{ij} \quad , \quad (\text{A6})$$

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

$$1 = \int p(C_{ij}|j, S, \alpha) dC_{ij} \quad . \quad (\text{A7})$$

Here we have also introduced some “hyperparameters”  $\alpha$ , which are variables which parameterize prior distributions. The subset of hyperparameters  $\alpha$  which apply to  $p(C_{ij}|j, S, \alpha)$  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  $p(C_{ij}|j, S, \alpha)$  varies slowly around the best-fit amplitude,

$$p(\mathbf{F}_i|j, S, \alpha) \propto \exp\left(-\frac{1}{2} \tilde{\chi}^2\right) p(\tilde{C}_{ij}|j, S, \alpha) \sigma_{C_{ij}} \quad , \quad (\text{A8})$$

where  $\tilde{\chi}^2$  is the best-fit chi-squared,  $\tilde{C}_{ij}$  is the best-fit amplitude, and  $\sigma_{C_{ij}}$  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  $\sigma_{C_{ij}}$  of the best-fit amplitude.

Marginalization over the template space looks like

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

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

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

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

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

$$\begin{aligned} p(\mathbf{F}_i|C_{ij}, j, S) &\propto \exp\left(-\frac{1}{2} \chi^2\right) \\ p(\mathbf{F}_i|j, S, \alpha) &= \int p(\mathbf{F}_i|C_{ij}, j, S) p(C_{ij}|j, S, \alpha) dC_{ij} \\ p(\mathbf{F}_i|S, \alpha) &= \sum_j p(\mathbf{F}_i|j, S, \alpha) P(j|S, \alpha) \quad . \end{aligned} \quad (\text{A11})$$

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

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

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

This model is fully generative; it specifies for any observed flux  $\mathbf{F}_i$  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  $i$ :

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

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

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

The likelihood  $p(\{\mathbf{F}_i\}|\boldsymbol{\alpha})$  is the total, marginalized likelihood for the combined data set of all the observations  $\mathbf{F}_i$  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  $p(\{\mathbf{F}_i\}|\boldsymbol{\alpha})$  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)  $\Omega_i$

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

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.