Separating Stars and Galaxies Based on Color

Victoria Strait
Furman University, 3300 Poinsett Hwy, Greenville, SC 29613 and
University of California, Davis, 1 Shields Ave., Davis, CA, 95616

Using photometric data from the Deep Lens Survey (DLS) we develop a star-galaxy separation algorithm based on objects' colors in six bands (B,V,R,z,J,K). Using a training set selected from a catalog of stars classified via their DLS shapes, we fit a third order polynomial to approximate the stellar locus. Our algorithm produces a weighted probability of an object being a star. Based on each object's distance from the stellar locus in color space, we fit the resulting histogram as the sum of two Gaussians. We find that near-infrared information (J and K) provide the best separation, but explore using optical information alone to determine the classification as well. Our results demonstrate that the use of color information has the potential to dramatically improve star-galaxy classification when used in conjunction with existing shape-based algorithms.

I. INTRODUCTION

The advent of large telescopes in the modern era of astronomy have pushed the amount of data available to new levels. However, as we push to study fainter galaxies at higher redshifts, we are still limited by the resources available. One of the most fundamental problems is that of object classification, as we must first isolate the objects that we wish to study amongst the millions of objects observed. One of the most basic and most important classifications is that of star/galaxy separation, that is, separating the population of stars local to our Milky Way galaxy from external galaxies. Studies of Galactic structure need a clean sample of local stars to identify the structure of Galactic components without contamination from faint galaxies, which could bias estimates of these structures. Cosmological studies, on the other hand, often rely on the large-scale clustering of galaxies to determine the growth of structure, so a clean sample with no local stars is optimal for clustering measurements. However, more importantly, one of the most powerful tools for modern cosmology is the use of weak gravitational lensing, where measurement of the distortion of galaxy shapes by intervening massive structure is used to infer the amount of mass along the line of sight to distant galaxies. In order to calibrate this very small effect, we must understand the imprint of the telescope and detector on the measured shapes of objects. This is usually done by identifying a set of point-like reference stars in order to measure the point-spread-function (psf) induced by the atmosphere and detector. Thus, identifying a pristine sample of stars is essential for weak lensing measurements.

The information available for each object consists of both its measured shape, as well as information about the object's spectral energy distribution (SED), the amount of flux the object emits as a function of wavelength. Nearly all star/galaxy classifiers to date[1,4] use shape alone to determine whether an object is a point-like star, or an extended galaxy. In this paper, we will examine using an object's color to determine the likelihood that it is a star or a galaxy.

If a high resolution spectrum is available for an object, star/galaxy classification is almost always trivial, as spectral features unique to stars or galaxies are easily identifiable. However, due to limitations in telescope time and current detector technology, we do not have a spectrum for each object, rather we only have total flux measured through a broad band filter. The set of filters can be thought of as a very low resolution spectrum. We define an object's color as the (log) ratio of the fluxes in two filters. Due to the limited number of such filters, ambiguities between object types appear, as objects with very different spectra can have very similar colors.

The Deep Lens Survey is a twenty square degree, ground-based, multi-band survey which provides information about local and deep-space objects.[5] Specifically, it provides flux values for objects (galaxies, stars, quasars, etc.) from which spectra can be plotted and color can be calculated. As seen in figure 1, the spectrum of an object is a plot of its wavelength vs. flux; the spectra of objects are uniquely shaped, which is the key to being able to use flux values to decipher them.

Filters (or bands), as related to spectra, are pieces of glass that allow only certain ranges of wavelengths through, so that an object can be measured from the viewpoint of several bands separately and compared. These bands are useful when calculating color for an object, which is defined using two fluxes from the spectrum of one object. The equation for a color index is as follows:

Color X-Y=-2.5log(Fx/Fy) where Fx is flux of x, Fy is flux of y

It is common to calculate multiple color indices for an object, and plot one against the other; this creates a color-color diagram. Color-color diagrams are the chief piece of information that we will use for separating objects based on their color. Some color-color diagrams show us visually the separation between stars and galaxies, as seen on the right in figure 2, while others show very little separation, as seen on the left in figure 2. In both cases, there exists a stellar locus; this is defined as the imaginary line where stars fall on a color-color diagram. The line is somewhat straight, but imperfections

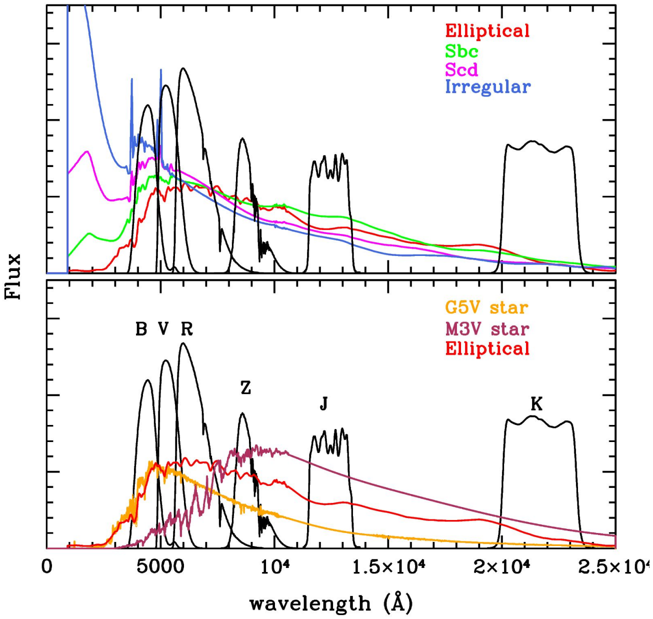

FIG. 1. Four galaxies' spectra are shown on the top panel: an elliptical, irregular, and two spiral, with two stars and an elliptical galaxy shown on the bottom panel. In black, 6 filters are plotted to show the ranges of wavelengths which they cover. They are labeled on the bottom panel, B, V, R, z, J and K. Each spectrum has a unique shape, and therefore a unique value in each filter. Using the ratio of values in two separate filters offers a more unique definition of an object's color. Seen later in the formal definition of color, multiple filters are used to identify objects more accurately.

are evident more so in some color-color diagrams than in others. In either case, the stellar locus can be defined by fitting exclusively stellar data to a polynomial, and then applying that line to the entirety of the data (including non-stellar objects).

Color-color diagrams become more complicated when the objects shown are at fainter magnitudes. Because galaxies can be seen to great distances that redshift their spectra, they have lines that serve as functions of where they show up on color-color diagrams based on how far away from the observer they are, while stars remain in very similar regions of the color-color diagrams because they are all local to our own Galaxy. This comes across as galaxies showing up all over the diagram, including along the stellar locus as objects become fainter. The stellar locus is better defined at brighter magnitudes.

The general problem we are trying to answer is whether there is a good way to utilize color-color diagrams to separate star and galaxy populations. The general method, described further below, is to find the stellar locus, and then find distances of each object from that line. We will then have a population of objects that is close to the stellar locus, and a population that is farther away (stars and galaxies, respectively).

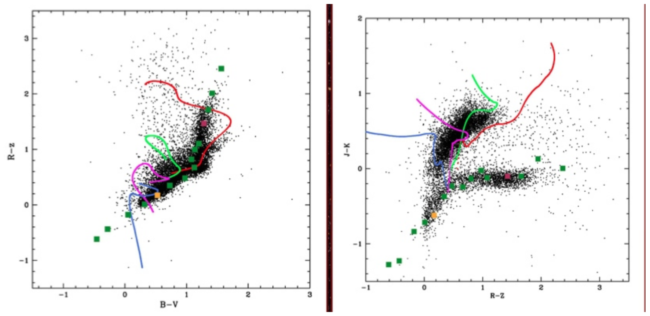

FIG. 2. Color-color diagrams are flux ratios plotted against each other in different combinations. They force populations to show up differently, depending on the ratios (color indices) used. In the above figure it can be seen on the right side that R-z vs. J-K, populations of stars and galaxies are very well separated, while on the left side, B-V vs. R-z shows populations that are less well separated, and instead fall somewhat in the same place on the diagram.

II. METHODS/DATA

From the Deep Lens Survey, we are provided with fluxes of several thousand objects in various filters; choosing the color indices in which to work is the first task. It was previously known that colors in the near-infrared, specifically R-z vs. J-K provides well-separated populations, making a good starting point for a star-galaxy separator. The next preliminary step was to decide on a magnitude at which to cut objects, so that there was still some separation to work with. We chose R = 20, as there was enough data to use to fit a polynomial for the stellar locus, but there was also still an obvious separation.

Next, we queried a DLS database called StarDB, which is a collection of exclusively stellar objects. These were identified using a shape-based algorithm, and were hand-verified as being accurately classified as stars. There were some objects that fell into a population other than the stellar locus, that was discovered in Covey et al. to be quasars. Subtracting these objects from the database, and taking the data of stars with a magnitude of greater than 18 (R ¿ 18), we used the downhill simplex method to fit the data to a polynomial.[3]

The downhill simplex method of fitting data to a polynomial works as follows: A simplex, which is defined as a shape of N+1 vertices in an N-dimensional space (a triangle for our 2-dimensional space), is created with data points. The shape is then reflected, contracted, and expanded with other data points, finding fixed minima values where data are most highly concentrated. This is done repeatedly until the data are fit to a polynomial. The algorithm can fall into local minima if not enough range is specified.

After fitting the stellar data with a polynomial (now, the stellar locus) and overlaying it onto the entirety of the data, we binned the data using 12 lines tangent to the polynomial, in which the bin sizes are equal in the x direction, as seen in figure 3. This led to some discre-

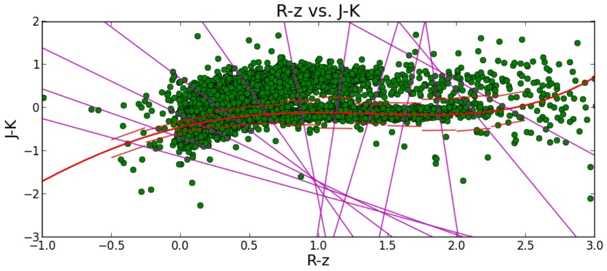

FIG. 3. This is all bright data, including galaxies. We have placed 12 bins that are orthogonal to the stellar locus. The plot is distorted, so visually the lines are not obviously orthogonal.

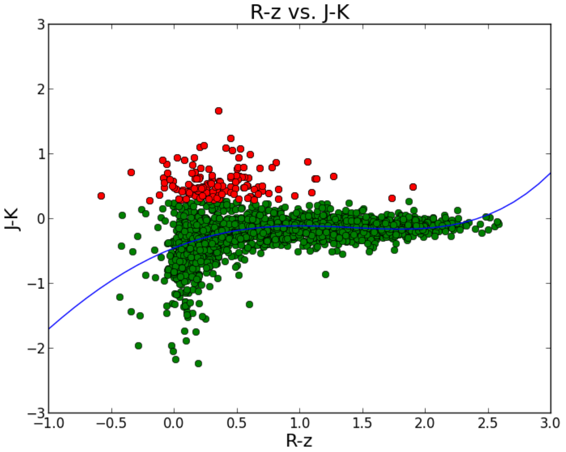

FIG. 4. The red data points in the figure are not stars, and in Covey et al it was found that in J-K at above a value of .27, a cut can be made to assume that any object above this line is a quasar. These objects made it into StarDB because of their point-like shape. Disregarding the quasars, we used a minimizer to fit the data to a polynomial, which defines the stellar locus. This entailed defining a likelihood equation, and using the downhill simplex method.

tion when the bins were tilted, because transformations needed to be made. This method captured the majority of the data points in some bin, but for those that it didn't, a stipulation was added that if an object was not in a bin, its distance to all the bin centers was calculated, and it was placed in the bin which had the closest center.

For the data in each bin, we found the Euclidian distance from the point to the stellar locus and binned them in a histogram. We then fit the histograms as a sum of two Gaussians; the left distribution with mu set at zero, and the other mu and sigmas set at initial guesses and minimized to fit the data, as seen in figure 4. Using the likelihood equation, a probability is then assigned to each object based on where in the curve it is. This is only a problem if an object falls into the part of the curves where there is an overlap of the two gaussian fits. Otherwise, the probability is 0 or 1. The likelihood equation is defined as:

$$f(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

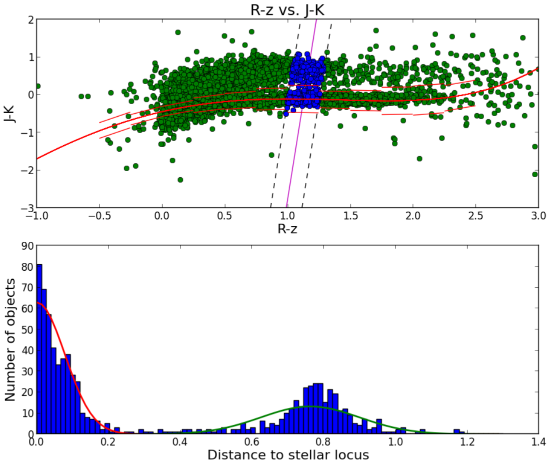

FIG. 5. Shown above is a bin in the middle of the data, which is the best case scenario. Here, data is completely separated into two populations: stars under the half-gaussian on the left, and galaxies under the curve on the right. There is visible separation even in the color-color diagram, and there are no ambiguous objects. Shown on the bottom are the two gaussian distributions corresponding to stars on the left and galaxies on the right.

III. RESULTS

The probability that the object is a star is based on the values of the two Gaussians at the distance of the object in the bin by:

P(star) = P(star gaussian(dist))/ (P(star gaussian(dist)) + P(galaxy gaussian(dist))

In the best case scenario, the population of stars is completely separated from the population of galaxies, such as in figure 5. There is whitespace visible even in the color-color diagram, and there is a visible separation in the gaussian curves. In these cases, the objects that fall under the half-gaussian can be identified as stars, and those that fall under the curve on the right can be identified as galaxies.

When the populations overlap, results become somewhat more complicated. Such as in figure 6, where there is a small amount of overlap in the populations, where these objects must be assigned a probability that is not 0 or 1.

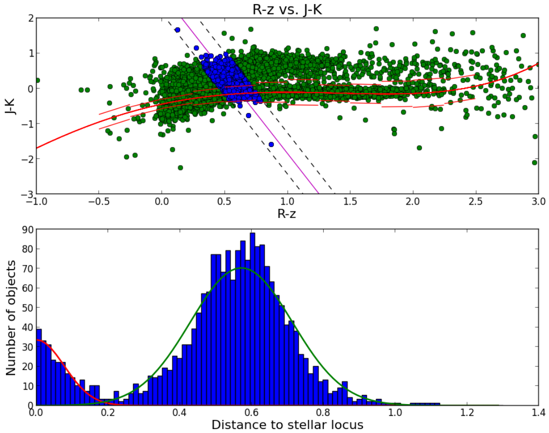

FIG. 6. Shown above is a bin located slightly to the left of center. Here the curves begin to overlap, creating some ambiguity which requires a normalized probability in order to assign each object located under the overlap with a likelihood. Shown below are the gaussian distributions, corresponding to stars on the left and galaxies on the right.

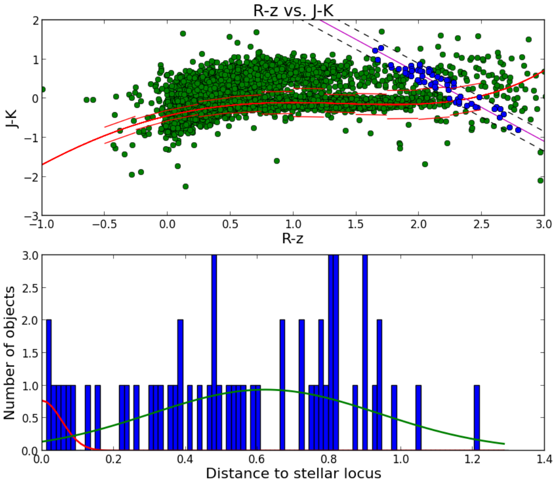

FIG. 7. Shown above is a bin located on the far side of the data. Where objects are scarce, either because we have chosen only bright objects, or simply because of the lack of objects of that color, the algorithm fails expectedly. This is simply because of small number statistics, and cannot be helped. Shown below is the histogram and two gaussian curves corresponding to stars on the left and galaxies on the right.

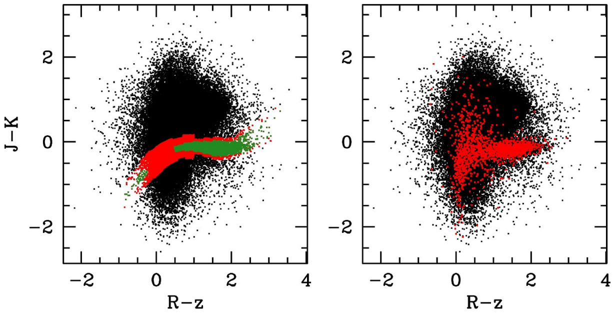

FIG. 8. Shown on the left are two probability cuts, where P¿.90 in the middle of the data. P¿.50 along the stellar locus. On the right, the results of the dlsqc shape algorithm are shown where stars appear in red and galaxies in black.

In the worst case scenario, such as on either far side of the data in figure 7, there is a scarce amount of data, and fitting it to a curve is unsuccessful. This is expected for any bin where there are very few data points.

All of the results discussed above were from only the R-z and J-K color indices; the algorithm worked well for those near infrared bands, but when it was tested on B-V vs. R-z (where there is little separation between populations in color-color diagrams), the results were not as satisfying. This is because, as in figure 8, the algorithm which is based on distance from the stellar locus does not perform as well when most of the objects are clustered around the stellar locus. There is some separation towards the ends of the data; unfortunately this is also where we encountered problems with small number statistics before.

To show our final results, we plotted objects with two probability cuts: P is greater than .50, and P is greater than .90, as can be seen in figure 8. There are more data than seen previously in this plot because we added in all data, including those objects with faint magnitudes. This produces a plot with more galaxies in places where we have only seen stars so far, such as in the lower right hand side of the plot. (This can be attributed to redshift tracks, as discussed previously.)

IV. CONCLUSIONS AND FUTURE WORK

The algorithm's results are what was expected. A cluster of objects around the stellar locus identifying as stars, and ones closer to the line with a higher probability of being a star. There is an issue with the actual stellar locus polynomial fit, which is that it is slightly misshapen in that it should bend more toward zero on the left. This can be attributed to a lack of blue stars in the original StarDB data; this must have caused the fit to be pulled upward instead of following the majority of the data.

Future work entails fixing the polynomial fit by adding test stars to the StarDB data, and perfecting the algorithm in R-z vs J-K. Following that, we will implement this algorithm or another on several other color-color combinations. The ultimate goal of the project is to have multiple algorithms for color-color diagrams, weight them based on importance and the amount of information given, and finally combine them with the shape algorithm for a superior separator.

In figure 9, the results of the shape algorithm are shown as galaxies on the left and stars on the right. There are problems with both; the galaxy results picked up some objects on the stellar locus, probably because they were stars with a less point-like shape because of their distance. The star results included objects which we know are not stars but quasars. Our color-based algorithm combined with the shape algorithm will eventually correct for both of these problems. Using the information from both shape and color will provide a superior classification to an algorithm that uses only one of the two pieces of information.

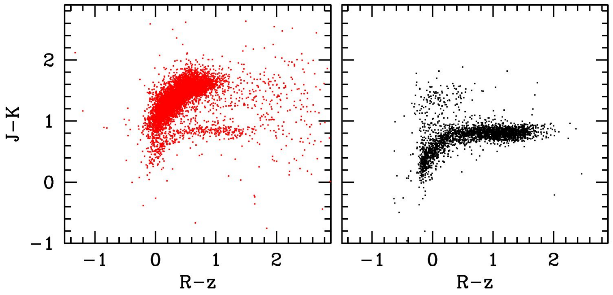

FIG. 9. Shown on the left are stars detected by the dlsqc shape algorithm, and galaxies on the right. There are objects in both plots that have been misidentified: objects along the stellar locus in the left plot, and objects above the stellar locus (quasars) on the right plot.

[1] M. T. Sougmac et. al *Star/galaxy separation at faint magnitudes: Application to a simulated Dark Energy Survey 12* 2012: Mon. Not. R. Astron. Soc.

[2] Ross Fadely, David W. Hogg Beth Willman *Star-galaxy Classification in Multi-Band Optical Imaging.3.12* 2012: apsto-ph

[3] Press, William H., Teukolsky, Saul A., Vetterling, William T. and Flannery, Brian P. *Numerical Recipes in C (@nd Ed.): The Art of Scientific Computing* 1992: Cambridge University Press, New York, NY, USA

[4] Mamon, G.A., Borsenberger, J., Tricottett, M. and Banchet, V. *Galaxies with DENIS: Preliminary star/galaxy separation and first results* 1998: Kluwer, ASSL, vol. 230

[5] *Deep Lens Survey Website* 2014, http://matilda.physics.ucdavis.edu/working/website/index.html