STAR/GALAXY SEPARATION IN HYPER
SUPRIME-CAM AND MAPPING THE MILKY WAY WITH STAR COUNTS

JOSÉ ANTONIO GARMILLA

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF
ASTROPHYSICAL SCIENCES
ADVISER: MICHAEL STRAUSS & ROBERT LUPTON

SEPTEMBER 2016
© Copyright by José Antonio Garmilla, 2016.

All rights reserved.
Abstract

We study the problem of separating stars and galaxies in the Hyper Suprime-Cam (HSC) multi-band imaging data at high galactic latitudes. We show that the current separation technique implemented in the HSC pipeline is unable to produce samples of stars with $i \gtrsim 24$ without a significant contamination from galaxies ($\gtrsim 50\%$). We study various methods for measuring extendedness in HSC with simulated and real data and find that there are a number of available techniques that give nearly optimal results; the extendedness measure HSC is currently using is among these. We develop a star/galaxy separation method for HSC based on the Extreme Deconvolution (XD) algorithm that uses colors and extendedness simultaneously, and show that with it we can generate samples of faint stars keeping contamination from galaxies under control to $i \leq 25$. We apply our star/galaxy separation method to carry out a preliminary study of the structure of the Milky Way (MW) with main sequence (MS) stars using photometric parallax relations derived for the HSC photometric system. We show that it will be possible to generate a tomography of the MW stellar halo to galactocentric radii $\sim 100$ kpc with $\sim 10^6$ MS stars in the HSC Wide layer once the survey has been completed. We report two potential detections of the Sagittarius tidal stream with MS stars in the XMM and GAMA15 fields at $\approx 20$ kpc and $\approx 40$ kpc respectively.
Acknowledgements

First and foremost, I thank my advisors Prof. Michael Strauss and Prof. Robert Lupton. This thesis would not have been possible without their dedicated support. Both of them were as generous as circumstances allowed with the one thing they can’t get back—their time. I will always be grateful to them for that. Furthermore, I had the great fortune to be advised by people whose knowledge on areas relevant to the problems tackled here is as developed as a human mind can possibly achieve.

Prof. Michael Strauss is an exceptionally well-rounded scientist. He has carried out outstanding work in a wide range of topics in Astronomy, and he is familiar with both the scientific and the technical details of the research in these areas. He has done an impressive amount of reading on recent developments in Astronomy, and this gives him a very clear vision of where things are going. Often times, after having a discussion with him, he would point me to a handful of related publications to complement our exchange. This turned out to be extremely helpful, and it substantially sped up the process of determining what to do and to find the place for this thesis in the much larger body of related work produced by the community.

For the past three years I had the rare privilege of regularly getting the attention of one of the world experts on the reduction of optical data in Astronomy, Prof. Robert Lupton. He routinely provided me with ancillary tools to be able to do what I needed to do. I greatly benefited from his extensive knowledge on CCDs, Astronomy, Telescopes’ Optics, Software, Statistics, Numerical Methods, Programming Languages and many other things that not only made my day-to-day endeavors easier, but even placed what would have been intractable goals within reach.

Besides my advisors, this thesis was also made possible in no small part by Dr. James Bosch. A great deal of the work in this thesis relies heavily on two software tools for which he has played a key role in the development: the Hyper Suprime-Cam (HSC) pipeline, and GalSim. Ever since I started using these tools, Dr. Bosch
has made himself amply available to answer my questions, help me solve problems related to these tools, and he has always given serious consideration to my requests for additions and modifications. Other members of the HSC software team occasionally assisted me, and I’d like to thank them for that: Dr. Paul Price, Prof. Naoki Yasuda, Prof. Masayuki Tanaka, Dr. Hironao Miyatake, and Dr. Bob Armstrong.

During the period in which I worked on this thesis, I had yearly meetings with a committee composed of members of the Astrophysics department that made comments on my progress and gave me advice on how to proceed. The committee was composed of: Prof. Michael Strauss, Prof. Robert Lupton, Dr. James Bosch, Prof. David Spergel, and Prof. James Gunn. I found these meetings particularly useful; each of these meetings truly marked a turning point in the direction towards which my efforts aimed. I want to specially thank Prof. Michael Strauss and Prof. David Spergel who also served as readers in my defense committee, and supplied me with invaluable comments and corrections after reading early manuscripts of this thesis. I’d also like to thank in advance those who will serve as examiners in my defense committee: Prof. Michael Strauss, Prof. Robert Lupton, and Prof. Željko Ivezić.

I had extensive discussions with Dr. Adrian Price-Whelan, who is an expert on stellar streams and other structures in the stellar halo. Without these discussions I would not have been able to interpret the results I report in §5.3. He suggested that the excess counts of stars we were seeing were due to the Sagittarius stream, and he pointed me to the Law and Majewski [2010] simulation to confirm that. He also helped me assess the potential scientific impact of a more extensive study of the stellar halo with HSC.

I’m grateful to the HSC collaboration for entrusting me with their data and resources, and for the generous financial support that allowed me to concentrate on this work. The HSC collaboration includes the astronomical communities of Japan and Taiwan, and Princeton University. The HSC instrumentation and software were
built by the National Astronomical Observatory of Japan (NAOJ), the Kavli Institute for the Physics and Mathematics of the Universe (Kavli IPMU), the University of Tokyo, the High Energy Accelerator Research Organization (KEK), and the Academia Sinica Institute for Astronomy and Astrophysics in Taiwan (ASIAA). Funding was contributed by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), the Japan Society for the Promotion of Science (JSPS), NAOJ, Kavli IPMU, KEK, ASIAA, the Toray Science Foundation and Princeton University.

I’d also like to thank the many people that I haven’t mentioned and I interacted with while at the Princeton Graduate Program. I recognize that during my time at Princeton I profited from the remarkable community of students, postdocs, and faculty. I had too many conversations to remember from which I learned something that I would probably not have found on my own, and was of immediate relevance to my work. From these interactions, I also learned many valuable things about other areas of research in Astronomy and other disciplines that have enhanced my technical capabilities and enriched my thinking. Among these people are: Roman Rafikov, Caleb Bastian, Cristobal Petrovich, Jason Li, Simone Ferraro, Peter Melchior, Emmanuel Schaan, James Stone, Yan-Fei Yiang, and Timothy Brandt.
Preface

The idea for this thesis came into being during a conversation with Prof. Robert Lupton about projects for a thesis. He mentioned the Hyper Suprime-Cam (HSC), and some of the challenges he foresaw because of the expected quality of the data. Among the things he mentioned was the need to improve on star/galaxy separation: due to the much larger number of marginally resolved galaxies deeper HSC catalogs would contain, Prof. Lupton didn’t expect that the approach he used in the Sloan Digital Sky Survey would suffice. This stroke me as an important and tractable problem, and so I decided to make it my thesis project.

Inspired by the work of Fadely et al. [2012], we initially hoped to derive a classifier using spectral templates matched to HSC colors that would yield clean samples of stars and galaxies up to HSC’s limiting magnitudes. This plan was quickly thwarted by persistently dismal results that led us to change to a more empirical approach—Supervised Learning. Prof. Alexie Leauthaud, a member of the HSC collaboration, had provided an updated version of the HST/ACS catalog of the COSMOS field she published in Leauthaud et al. [2007] for a variety of tests. Prof. Lupton suggested that I use this catalog to generate a truth table for supervised learning. This turned out to be a much more profitable method, and we started to see promising results with several supervised learning algorithms early on.

Despite the encouraging performance of supervised learning techniques, there remained a concern. Fadely et al. [2012] give a stern warning about using supervised learning classifiers in datasets with lower S/N than that of the data they were trained with, and we believe rightly so. We were however worried about a more general problem; what’s known as sampling bias in the Machine Learning literature—using a classifier on data that has a significantly different probability distribution from the data that trained the classifier. In photometry, this can be caused by a variety of things besides differences in S/N: different seeing, sky brightness, crowding and so on.
We were estimating the performances in the truth table itself, and we didn’t know how well these results would generalize to datasets with a wider range of conditions.

We eventually ran into the Extreme Deconvolution (XD) paper by Bovy et al. [2011b], and we realized that XD could address our sampling bias concerns: XD treats the error bars for what they are instead of just another set of inputs. The hope was that there would be enough information about the observing conditions in the error bars for XD to counteract the bias in datasets with significant sampling bias. We tested this by using a set of shallower truth tables with reductions of data with different seeing conditions produced by Prof. Masayuki Tanaka, and we confirmed that XD indeed adapts itself to different observing conditions and produces acceptable results even when the training data is deeper and has better seeing than the test data. Finally, we were satisfied with our star/galaxy classifier, and we began to work on developing the tools for a tomography of the Milky Way stellar halo in the manner of Jurić et al. [2008] with HSC.

All the work in this thesis was carried out by me with few exceptions that the text clearly notes and gives due credit to the sources. None of its results have been published in a peer reviewed journal; I’m still in the process of submitting parts of this thesis for peer review. However, some of the methods and results discussed here are being used for future publications by the HSC collaboration, one of which reports the discovery of a new Milky Way satellite and has been submitted to The Astrophysical Journal: Homma et al. 2016, A New Milky Way Satellite Discovered in the Subaru/Hyper Suprime-Cam Survey.

The thesis is organized in five chapters. Chapter 1 is an introduction to the star/galaxy separation problem: we review recent efforts to improve on what is traditionally done, and we put the work in this thesis on that context. In Chapter 2 we describe how we generate a truth table that we use to train and test star/galaxy separation techniques. In Chapter 3 we study the problem of measuring the extendedness
of a source, and using it for star/galaxy separation with both simulated images and HSC data. In Chapter 4 we make use of HSC colors for star/galaxy separation, and finally we combine colors with extendedness using the XD method. In Chapter 5 we use our separator that combines colors and extendedness and apply it to 100 deg$^2$ of HSC Wide layer data available in January of 2016 to forecast the scientific payoff of a study of the MW stellar halo in the full HSC Wide layer, and we report a tentative detection of two sections of the Sagittarius stream.
To the memory of José Salvador.
Contents

Abstract ................................................................. iii
Acknowledgements ....................................................... iv
Preface ................................................................. vii
List of Figures ......................................................... xiii

1 Introduction .......................................................... 1

2 Building a Truth Table ............................................... 7
   2.1 Matching Catalogs ............................................. 11
   2.2 Testing Labels .................................................. 14

3 Measuring Extendedness ............................................. 21
   3.1 Extendedness Measurement in the HSC Pipeline .......... 22
   3.2 Simulating Extendedness Measurements .................... 29
   3.3 Extendedness Measurements on HSC Data .................. 60

4 Using HSC Colors .................................................... 66
   4.1 Using Colors Only ............................................. 66
   4.2 Combining Colors and Extendedness ......................... 79
   4.3 Comparison to Support Vector Machines .................... 86
   4.4 Changes in Seeing Conditions ................................. 90
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Leauthaud et al. 2007 star/galaxy separation</td>
<td>9</td>
</tr>
<tr>
<td>2.2</td>
<td>MAG_AUTO vs. CModel Magnitude in HSC-I for COSMOS objects</td>
<td>13</td>
</tr>
<tr>
<td>2.3</td>
<td>Extendedness vs. CModel Magnitude diagram</td>
<td>17</td>
</tr>
<tr>
<td>2.4</td>
<td>Color-color diagrams in HSC-R, HSC-I, and HSC-Z</td>
<td>18</td>
</tr>
<tr>
<td>2.5</td>
<td>Same as figure (2.4) but in HSC-G, HSC-R, and HSC-I</td>
<td>19</td>
</tr>
<tr>
<td>2.6</td>
<td>Same as figure (2.4) but in HSC-I, HSC-Z, and HSC-Y</td>
<td>20</td>
</tr>
<tr>
<td>3.1</td>
<td>Mag_{psf} − Mag_{cmodel} in HSC-I histograms for the objects in figure (2.3)</td>
<td>25</td>
</tr>
<tr>
<td>3.2</td>
<td>Mag_{psf} − Mag_{cmodel} cuts in HSC-I</td>
<td>27</td>
</tr>
<tr>
<td>3.3</td>
<td>Purity and completeness for cuts on Extendedness</td>
<td>28</td>
</tr>
<tr>
<td>3.4</td>
<td>Calibration of simulation’s background noise using real HSC data</td>
<td>30</td>
</tr>
<tr>
<td>3.5</td>
<td>Illustration of extendedness measures</td>
<td>36</td>
</tr>
<tr>
<td>3.6</td>
<td>Histograms of double gaussian parameters in HSC-G wide</td>
<td>37</td>
</tr>
<tr>
<td>3.7</td>
<td>Histograms of double gaussian parameters in HSC-R wide</td>
<td>38</td>
</tr>
<tr>
<td>3.8</td>
<td>Histograms of double gaussian parameters in HSC-I wide</td>
<td>39</td>
</tr>
<tr>
<td>3.9</td>
<td>Histograms of double gaussian parameters in HSC-Z wide</td>
<td>40</td>
</tr>
<tr>
<td>3.10</td>
<td>Histograms of double gaussian parameters in HSC-Y wide</td>
<td>41</td>
</tr>
<tr>
<td>3.11</td>
<td>Comparison of extendedness measures without noise</td>
<td>42</td>
</tr>
<tr>
<td>3.12</td>
<td>Same as figure (3.11) but different PSFs</td>
<td>43</td>
</tr>
<tr>
<td>3.13</td>
<td>Comparison of extendedness measures with noise</td>
<td>46</td>
</tr>
<tr>
<td>3.14</td>
<td>Same as figure (3.13) but using circular de Vaucouleurs profiles</td>
<td>47</td>
</tr>
<tr>
<td>Section Number</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>----------------</td>
<td>------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>3.15</td>
<td>Extendedness sensitivity to noise as a function of input radius</td>
<td>48</td>
</tr>
<tr>
<td>3.16</td>
<td>Extendedness measures vs Magnitude in simulations</td>
<td>53</td>
</tr>
<tr>
<td>3.17</td>
<td>Extendedness histograms in wide depth simulations</td>
<td>54</td>
</tr>
<tr>
<td>3.18</td>
<td>Scores for cuts in extendedness measures</td>
<td>55</td>
</tr>
<tr>
<td>3.19</td>
<td>Extendedness histograms in single epoch simulations</td>
<td>56</td>
</tr>
<tr>
<td>3.20</td>
<td>True vs noisy extendedness measurements in simulations</td>
<td>57</td>
</tr>
<tr>
<td>3.21</td>
<td>Correlations among extendedness measures without noise</td>
<td>58</td>
</tr>
<tr>
<td>3.22</td>
<td>Correlations among extendedness measures with noise</td>
<td>59</td>
</tr>
<tr>
<td>3.23</td>
<td>Same as figure (3.16) but with HSC-I data</td>
<td>61</td>
</tr>
<tr>
<td>3.24</td>
<td>Same as figure (3.17) but with HSC-I data</td>
<td>62</td>
</tr>
<tr>
<td>3.25</td>
<td>Scores for $\Delta$Mag cuts in HSC-I full depth data</td>
<td>63</td>
</tr>
<tr>
<td>3.26</td>
<td>Scores for $r_{tr} - (r_{tr})_{PSF}$ cuts in HSC-I full depth data</td>
<td>64</td>
</tr>
<tr>
<td>3.27</td>
<td>Scores for $\Delta$Mag cuts in HSC-I wide depth data</td>
<td>64</td>
</tr>
<tr>
<td>3.28</td>
<td>Scores for $r_{tr} - (r_{tr})_{PSF}$ cuts in HSC-I wide depth data</td>
<td>65</td>
</tr>
<tr>
<td>4.1</td>
<td>XD fit ellipses in bin $18 &lt; \text{Mag}_{cmodel} \text{HSC} - I &lt; 22$</td>
<td>69</td>
</tr>
<tr>
<td>4.2</td>
<td>XD fit ellipses in bin $22 &lt; \text{Mag}_{cmodel} \text{HSC} - I &lt; 24$</td>
<td>70</td>
</tr>
<tr>
<td>4.3</td>
<td>XD fit ellipses in bin $24 &lt; \text{Mag}_{cmodel} \text{HSC} - I &lt; 25$</td>
<td>70</td>
</tr>
<tr>
<td>4.4</td>
<td>XD fit ellipses in bin $25 &lt; \text{Mag}_{cmodel} \text{HSC} - I &lt; 26$</td>
<td>71</td>
</tr>
<tr>
<td>4.5</td>
<td>$P(\text{Star}</td>
<td>x, S_0)$ projected on color-color diagrams vs data</td>
</tr>
<tr>
<td>4.6</td>
<td>$P(\text{Star}</td>
<td>x, S_0)$ projected on color-color diagrams vs data</td>
</tr>
<tr>
<td>4.7</td>
<td>$P(\text{Star}</td>
<td>x, S_0)$ projected on color-color diagrams vs data</td>
</tr>
<tr>
<td>4.8</td>
<td>$P(\text{Star}</td>
<td>x, S_0)$ projected on color-color diagrams vs data</td>
</tr>
<tr>
<td>4.9</td>
<td>Visualize posteriors $P(\text{Star}</td>
<td>\text{Colors})$</td>
</tr>
<tr>
<td>4.10</td>
<td>$P(\text{Star}</td>
<td>\text{Colors})$ cuts scores</td>
</tr>
<tr>
<td>4.11</td>
<td>$P(\text{Star}</td>
<td>\text{Colors})$ cuts biases</td>
</tr>
<tr>
<td>4.12</td>
<td>Correlation between $\text{Mag}<em>{psf} - \text{Mag}</em>{cmodel}$</td>
<td>81</td>
</tr>
<tr>
<td>4.13</td>
<td>XD Gaussians in $r - i$ vs $\text{Mag}<em>{psf} - \text{Mag}</em>{cmodel}$ diagram</td>
<td>83</td>
</tr>
</tbody>
</table>
4.14 Data used to generate models in figure (4.13). ................................................. 84
4.15 Visualize posteriors $P(\text{Star}|\text{Colors + Extendedness})$. ........................................ 85
4.16 $P(\text{Star}|\text{Colors + Extendedness})$ cuts scores ..................................................... 85
4.17 $P(\text{Star}|\text{Colors + Extendedness})$ cuts biases ....................................................... 86
4.18 XD vs SVM. .................................................................................................................... 90
4.19 COSMOS Wide Seeing distributions in HSC-I. ............................................................. 91
4.20 COSMOS Wide truth tables in HSC-I. .......................................................................... 92
4.21 COSMOS Wide best seeing posteriors. ........................................................................ 92
4.22 COSMOS Wide median seeing posteriors. .................................................................. 93
4.23 COSMOS Wide worst seeing posteriors. ..................................................................... 93
4.24 COSMOS Wide best seeing scores .............................................................................. 94
4.25 COSMOS Wide median seeing scores ....................................................................... 94
4.26 COSMOS Wide worst seeing scores .......................................................................... 95
4.27 COSMOS Wide best seeing comparison to SVM. ......................................................... 96
4.28 COSMOS Wide median seeing comparison to SVM .................................................... 97
4.29 COSMOS Wide worst seeing comparison to SVM ....................................................... 97

5.1 Isochrones in LSST photometric system ........................................................................ 104
5.2 $r - i$ vs $r$ diagram with stellar posteriors. ................................................................... 105
5.3 Linear fit to stellar locus in $gri$ diagram. ........................................................................ 107
5.4 Linear fit to stellar locus in $riz$ diagram. ....................................................................... 107
5.5 Photometric parallax relations. ..................................................................................... 108
5.6 Tomography of COSMOS Field. .................................................................................... 109
5.7 HSC Wide in Galacthic coordinates. ............................................................................. 110
5.8 XMM Stars color-color diagrams. ................................................................................ 111
5.9 GAMA15 Stars color-color diagrams. .......................................................................... 112
5.10 Scores for star selector in radial and color bins. ............................................................ 114
5.11 HSC Wide star counts radial profiles. .......................................................................... 115
5.12 Color-magnitude diagram of stars in XMM field. .......................... 116
5.13 Same as figure (5.12) but in GAMA09 field. .............................. 117
5.14 Same as figure (5.12) but in WIDE12H field. .............................. 118
5.15 Same as figure (5.12) but in GAMA15 field. .............................. 119
5.16 Same as figure (5.12) but in HectoMap field. ............................ 120
5.17 Same as figure (5.12) but in VVDS field. ................................. 121
5.18 Same as figure (5.12) but in AEGIS field. ............................... 122
5.19 Right ascension and declination of stars in XMM field. ................. 123
5.20 Same as figure (5.19) but in the GAMA15 field. ........................ 124
5.21 Law and Majewski [2010] simulation. .................................... 126
5.22 Law and Majewski [2010] simulation in XMM and GAMA15. ........ 127

C.1 Multiband Extendedness comparisons. .................................... 143
Chapter 1

Introduction

The automatic separation of stars and galaxies is an old technical problem in optical astronomical imaging. Ever since digitized images became widespread, automatic star/galaxy separation has been an important step in the image processing tasks carried out by astronomers [e.g., MacGillivray et al., 1976, Sebok, 1979, Kron, 1980, Jarvis and Tyson, 1981, Valdes, 1982, Yee, 1991, Odewahn et al., 1992, Weir et al., 1995]. Later on, charge-coupled devices (CCDs) replaced photographic plates and star/galaxy separation techniques for the new technology started to appear in publications and software tools [e.g., Bertin and Arnouts, 1996, Bazell and Peng, 1998, Wolf et al., 2001, Lupton et al., 2001, Scranton et al., 2002, Ball et al., 2006, Gwyn, 2008, Henrion et al., 2011, Vasconcellos et al., 2011, Fadely et al., 2012, Soumagnac et al., 2013, Kim et al., 2015, Heinis et al., 2016]. After two decades of work on the subject new separation methods are still being proposed and there is no one best technique to do star/galaxy separation; instead, what we have is a variety of techniques that are still in use today. This diversity of approaches is in no small part due to different requirements that arise each time a star/galaxy separation procedure is needed. The choice of star/galaxy separation method is limited by the image processing framework used to do the measurements, the technical details of
the instrument that produces the data and the manner in which it’s used (i.e., seeing, depth, pixel scale, dithering strategy), and, maybe most importantly, the science goals. For example, some separation methods require an accurate model of the point spread function (PSF) to be estimated at each detected object’s position [e.g., Lupton et al., 2001, Soumagnac et al., 2013], and if the PSF is not well-sampled it may be impossible to build a reliable model of the PSF; in such a situation, a technique that requires no knowledge of the PSF may be a better choice [e.g., Bertin and Arnouts, 1996, Gwyn, 2008].

The separators that are used the most rely only on morphological information, and they have a simple approach: measure how extended an object’s image is—for example, by fitting a profile and looking at the best-fit radius—and based on that single number determine if it’s a star or a galaxy [e.g., MacGillivray et al., 1976, Sebok, 1979, Kron, 1980, Valdes, 1982, Yee, 1991, Lupton et al., 2001, Gwyn, 2008]. These techniques have been the most widely used for two good reasons: 1) they work well, and 2) they are easier to implement than the more sophisticated approaches. However, as we push existing datasets to lower signal-to-noise ratio (S/N) and acquire deeper datasets the first reason for these methods’ popularity fades. Pushing to lower S/N makes it harder to control for contamination of galaxies on samples of stars and vice versa, and deeper photometric surveys contain larger numbers of unresolved galaxies that are morphologically identical to stars. In addition, most photometric surveys focus on high galactic latitudes so that the number of galaxies per star grows rapidly with magnitude: this makes it impossible to extract clean samples of faint stars with any of these simple morphological cuts. Indeed, the inability to produce usable samples of faint stars in deep high-galactic-latitude images with morphological cuts has been the main motivation behind the recent efforts to develop more sophisticated star/galaxy separation techniques [Henrion et al., 2011, Vasconcellos et al., 2011, Fadely et al., 2012, Soumagnac et al., 2013, Kim et al., 2015, Heinis et al., 2016, and this thesis].
One of the most important issues addressed by recent publications on star/galaxy separation is the question of how to use information that is being neglected by the methods that make a cut in a proxy for extendedness. Clearly, there is information that should be useful and is simply not used: other morphological parameters, multiple extendedness measurements in multiband surveys, colors, and apparent magnitudes. A conceptually straightforward way of including more than one measurement in the separation is to use a Bayesian framework. This is the approach of Henrion et al. [2011] who combine morphological models’ likelihoods in different bands to compute a posterior probability of being a star. Similarly, Fadely et al. [2012] build a hierarchical Bayesian model with spectral energy distribution (SED) templates and the transmission curves of the photometric system to compute posterior probabilities based on colors alone. In this thesis we’ll follow the Bayesian approach as well to combine data from multiple bands.

Another aspect that recent publications have touched on is the use of machine learning (ML) algorithms. Even though this had already been explored previously [e.g., Odewahn et al., 1992, Weir et al., 1995], the increase in computing capabilities and the flurry of techniques and tools that has come out of the ML community in the past years justifies a revisit of ML applications to star/galaxy separation. Vasconcellos et al. [2011] proposed a method based on decision trees (DTs) that used over 10 morphological parameters, and was tested on Sloan Digital Sky Survey (SDSS) Data Release 7 (DR7) objects that had spectral classification. They were able to train a DT with equal or better performance than all other separation techniques used in SDSS before their work. Kim et al. [2015] were the first to use a ML technique called ensemble learning: this technique consists of combining multiple ML algorithms (usually a linear combination) to improve performance. They combined four separators: a simple morphological cut, self-organizing maps (SOMs), DTs, and the hierarchical Bayesian approach of Fadely et al. [2012]. They test their classifier on data from
CFHTLens \cite{Heymans2012, Erben2013}, and obtain higher performance than any of the four techniques would have given in isolation, demonstrating the potential benefits of ensemble learning. \cite{Heinis2016} used a combination of genetic algorithms (GAs) and support vector machines (SVMs) (see Appendix D for an introduction), where the GA is used to select the features the SVM will use for training from a long list of features: this is known as feature extraction. This method was demonstrated by the authors on Pan-STARRS1 data \cite{Kaiser2010} and they report promising results.

\cite{Soumagnac2013} discussed a more fundamental point: how to measure extendedness. In particular, they’ve proposed a new way of measuring extendedness that turned out to work well, and it’s been implemented in the popular image reduction package SExtractor \cite{Bertin1996}. In this thesis we also explore ways of measuring extendedness including that of \cite{Soumagnac2013} and compare them. We use simulated images and real data to study extendedness measures, and this has given us some insights into the effects of seeing, noise and the size and shape of the measured object. As we’ll see here, there’s reason to believe that the existing techniques to measure extendedness are very close to being optimal (see Chapter 3).

The goal of this work is not to propose another ML application to star/galaxy separation, or another novel way of measuring extendedness. The goal here is more specific, we want to produce a separation technique that works well for the Hyper Suprime-Cam (HSC) survey. The HSC survey is a large collaboration that includes researchers from the astronomical communities in Japan, Taiwan, and Princeton Uni-

\footnote{The idea behind genetic algorithms is inspired by the process of natural selection. A genetic algorithm starts with a set of candidate models that have to compete according to some rules. The models that do better get to produce versions of themselves with random mutations (offspring) for a subsequent round of competition with the next-generation models. The process repeats until some convergence criterion across generations is satisfied. Like in natural selection, the idea is that the “genes” that lead to the model that is best suited to win the competition (solve the problem) will eventually dominate over less helpful genes. In the case of \cite{Heinis2016} the genes are indicator functions that specify which features will be used for training with the SVM.}
The collaboration has been awarded 300 nights with the HSC instrument [Miyazaki et al., 2012], a 0.9-gigapixel wide-field camera in the prime-focus unit of the Subaru 8.2-m telescope on the summit of Maunakea in Hawaii. The camera is made up of 116 2K×4K Hamamatsu fully-depleted CCDs with a pixel scale of 0.17"/pixel and a field of view of 1.77 deg². The survey uses five broad-band filters ($g$, $r$, $i$, $z$, and $y$) and four narrow-band filters (N101, N387, N816, and N921). The full width at half-maximum (FWHM) of the PSF ranges between 0.4" and 1.0", and it is filter dependent. There are three layers in the survey: Wide, Deep, and Ultra-Deep. The Wide layer will cover 1,400 deg² with an integration time $\gtrsim 10$ min along the celestial equator with the five broad-band filters. The Deep layer will cover 27 deg² with an integration time $\gtrsim 1$ hrs with the five broad-band filters and three narrow-band filters (N387, N816, and N921). The Ultra-Deep layer will cover 3.5 deg² with an integration time $\gtrsim 10$ hrs with the five broad-band filters and three narrow-band filters (N101, N816, and N921). The $i$ band $5\sigma$ limiting magnitudes estimated with a 2" aperture for a point source are: 25.9 (Wide), 26.8 (Deep), and 27.4 (Ultra-Deep).

The survey is optimized for Weak gravitational lensing so it focuses on high galactic latitude fields, which puts it in the regime where galaxies outnumber stars at faint magnitudes. The current plan is to complete the survey by 2019. In this thesis we work with the January 2016 data release which includes $\approx 100$ deg² to full-depth in the Wide layer plus the Deep and Ultra-Deep fields with a fraction of the planned exposures.

We are particularly interested in giving HSC users a way of extracting clean samples of faint stars because this is the only limitation of the star/galaxy separation already in place in the HSC pipeline. The main motivation is to increase the scientific impact of HSC by enabling the study of the outer stellar halo of the MW with large numbers of stars. HSC has ground-breaking capabilities so a successful study of the stellar halo in the HSC survey is expected to lead to important findings—such a
study would discover a number of faint dwarf MW satellites and stellar streams, and
greatly improve our constraints on models for the MW stellar halo (see Chapter 5).
Chapter 2

Building a Truth Table

In this thesis we approach the problem of star/galaxy separation as a supervised learning problem. Appendix A has an introduction to supervised learning, and details on our methodology. Here we’ll just say briefly that solving a classification problem like star/galaxy separation as a supervised learning problem involves using a set of known stars and galaxies on which we have performed the same measurements we’ll perform on unlabeled objects that we want to classify into stars and galaxies. The set of known and labeled objects is also known as training set, or truth table. If the training set is large enough, and it’s representative of the dataset we want to predict labels for, it may be possible to use it to infer a classification rule that will perform better than random guessing on the predicted labels; in other words, statistical learning may be feasible [Vapnik, 2013]. Typically, when the two mentioned conditions are satisfied it’s possible to do learning with the truth table, and even when these conditions are violated learning has been shown to be possible. These are neither necessary nor sufficient conditions but are nevertheless important guidelines (see Appendix A).

Clearly, the very first thing to do in supervised learning is obtaining a truth table. For this, we use a public catalog from the Hubble Space Telescope (HST) COSMOS
survey [Leauthaud et al., 2007]. The COSMOS survey imaged a contiguous area of 1.64 deg$^2$ with the Advanced Camera for Surveys (ACS) Wide Field Channel (WFC) in HST to a 5σ limiting magnitude of F814W=26.5. The pixel scale of the ACS/WFC is 0.05", and the FWHM of the PSF is 0.12". The absence of an atmosphere allows for a small seeing (FWHM of the PSF) compared to ground-based cameras, and because of this star/galaxy classification can be done more reliably. The COSMOS data and catalogs are publicly available\footnote{http://irsa.ipac.caltech.edu/Missions/cosmos.html}, and HSC covers the COSMOS field completely in the Ultra Deep layer. In the January 2016 data release the Ultra Deep data in COSMOS already goes to $i > 27$ for most of the field.

Leauthaud et al. [2007] reduced the COSMOS data with the image processing package SExtractor [Bertin and Arnouts, 1996] and obtained a catalog of $1.2 \times 10^6$ objects. Stars and galaxies were separated in this catalog by looking at the ratio of peak surface brightness to total flux and choosing a hard cut by eye. SExtractor provides, among many other things, the measurements MU_MAX and MAG_AUTO. These are estimates of the peak surface brightness in magnitudes per square arc sec, and total apparent magnitude\footnote{In practice, any estimate of the total apparent magnitude won’t capture the total flux of the object in the image. MAG_AUTO is SExtractor’s best attempt to measure something that correlates well with the total flux of all objects in the image irrespective of their shape. This is typically achieved by using a window function (or aperture) that adapts to the shape of each object’s image. In the case of MAG_AUTO, an elliptical aperture is used where the ellipse is derived from the object’s image moments.} respectively. The left panel in figure (2.1) shows a plot of these two parameters for the objects in the catalog, and it shows two clearly distinct clusters of points. Since at a given MAG_AUTO the more concentrated an object is the lower MU_MAX, we expect stars to lie below galaxies. The classification simply draws a line that separates the two clusters by eye, and labels all the objects in the lower cluster stars and the objects in the upper cluster galaxies. The very few objects below the solid lines are more concentrated than the PSF, so they are dismissed as artifacts of the camera or cosmic rays. The figure also shows
that around MAG\_AUTO=25.5 the two clusters begin to merge, and because of this the authors drew a boundary for the box containing stars at MU\_MAX=21.5 between MAG\_AUTO=25.5 and MAG\_AUTO=26.5. This means that some stars with MAG\_AUTO ≥ 25.5 will be cut out of the box, and we have to be cautious when using labels from objects with MAG\_AUTO ≥ 25.5 in this catalog. The slope for the brightest stars is zero due to saturation, this does not concern us because we are interested in classification of faint objects. This separation procedure labels 3 × 10^4 objects in the catalog as stars.

Figure 2.1: Left Panel: Star/Galaxy classification in Leauthaud et al. [2007]. Stars are inside the solid lines. Galaxies are above the solid lines and the objects below it are artifacts of the camera or cosmic rays. MU\_MAX is the peak surface brightness estimate of SExtractor and MAG\_AUTO is its magnitude estimate. Right Panel: Apparent magnitude vs S/N for objects in this catalog (Figures from Leauthaud et al. [2007]).

The main reason the two clusters of points merge is that faint galaxies’ images are smaller. To see this, look at the trend the cluster associated with galaxies follows in the left panel of figure (2.1): the galaxies get closer to the stellar locus and eventually lie on top of it, marking the point at which galaxies are morphologically identical to stars in HST/ACS. There are two causes behind this trend. One is that the typical galaxy (R_e ∼ 5 kpc and M_V ≈ 21) subtends an angle ≤ 1” for z ≥ 0.5
which corresponds roughly to $i \approx 22$, and this is already getting close to the 0.12” FWHM of the HST/ACS PSF. At fainter magnitudes the typical galaxy will have a smaller image, so for faint enough magnitudes its image will be the PSF itself. The other reason behind the trend, is that we can only detect objects that have a peak surface brightness that is significantly higher than the background noise. If we consider an object with a fixed image size, say 1”, and move it along the MAG\_AUTO axis we’ll have to adjust it’s peak surface brightness to get the desired value for MAG\_AUTO. There will be a maximum value of MAG\_AUTO we can choose such that the peak surface brightness will be statistically distinguishable from a peak generated by background noise. Consequently, even if there are extended galaxies near the limiting magnitude they won’t be detected. In fact, the faintest detectable galaxies are necessarily point sources. The right panel of figure (2.1) shows S/N as a function of MAG\_AUTO for this catalog. We can see that at MAG\_AUTO=25.5 S/N $\sim 10$, so noise will cause scattering of galaxies into the box and of stars out of the box. We will ignore this and assume the labels are true.

We know that the star/galaxy classification in this catalog is not perfect, but as we will see it is more reliable than what we would get from an HSC catalog; even in the Ultra Deep layer which is considerably deeper than this catalog ($5\sigma$ limiting magnitude of $i = 27.4$ by the end of the survey), and it has five bands available (see §2.2). Thus, at least we can improve the quality of HSC star/galaxy separation by making it as close as possible to the classification we would obtain with the HST/ACS instrument. Furthermore, we’ll show that the limitations of these labels (i.e. objects with MAG\_AUTO $\gtrsim 25.5$ may not be reliably classified) are well beyond the limitations of any star/galaxy labels derived from the deepest HSC data. So even though the reliability of these labels at the faint end is an interesting and important question, it doesn’t affect us for our present goal which is to train classifiers with HSC data.
2.1 Matching Catalogs

To make use of the COSMOS catalog of the previous section, we have to match objects detected in HSC with objects in this catalog. We got the HSC Ultra Deep layer catalog of the COSMOS field generated with hscPipe 3.10.0 3 We matched them based only on their position in the sky and then looked at flux measurements in the two catalogs to double check the matches. The matching between the HST and HSC catalogs is done in two steps for each HST object. 1) Get the list of HSC objects that are inside a circle with a 1" radius centered at the HST object’s centroid. 2) If the list is not empty, match the HST object with the HSC object that is closest in that list; if the list is empty, that HST object won’t be matched to an HSC object. The matching across HSC bands is done by hscPipe 4 We match 606,720 objects across the two catalogs, and we find that 44% of the HST objects have no match in the HSC catalog: after visual inspection of a few cases we concluded that most of these are spurious detections in the HST/ACS catalog.

We noted a set of problematic objects in this match. These are objects labeled as stars in the HST/ACS catalog, while the HSC catalog reports them to be unambiguously extended (we discuss how extendedness is measured in HSC in chapter 3). Most of these problematic cases had to do with deblending problems in HSC 5 that is, HST/ACS was able to discern among multiple nearby objects while they appeared to be one blob to HSC because of its lower resolution. To eliminate these problematic objects in our list of matched objects, we plot the matched objects in a MAG\_AUTO vs. CModel magnitude 6 in HSC-I diagram in figure (2.2) and throw away outliers.

3The HSC processing pipeline is called hscPipe. It’s still being developed so the version numbers change frequently. See https://hsc-jira.astro.princeton.edu/jira/browse/HSC-1375.
4We are using forced photometry catalogs. This means that an effort is made to compute fluxes of the same object in a consistent way across HSC filters (same weight function, and centroid) so that the measured colors are meaningful.
5See https://hsc-jira.astro.princeton.edu/jira/browse/HSC-1212
6CModel magnitude is the hscPipe version of MAG\_AUTO: the best attempt at measuring something that correlates well with the total magnitude. In chapter 3 we describe how this measurement works.
In the two panels we draw the two lines that define the boundaries of the objects we admit in the match. The diagrams show that these problematic objects are a very small fraction of the total: these cuts only discard 14,926 objects leaving us with 591,794 entries in the truth table. We choose HSC-I because it is the closest to F814W in HSC. Since F814W and HSC-I have similar transmission curves and MAG_AUTO and CModel magnitude are estimates of the total magnitude in each filter, we expect the two magnitudes to be highly correlated and close to each other for objects that were correctly identified and measured in both catalogs. Problematic objects, like those that have deblending problems in HSC, should on the other hand have very different values for the two total magnitude estimates. There are also a few objects with an extreme value for CModel magnitude that do not appear on the diagram, these are simply discarded in our analysis.

The reason to plot stars and galaxies separately (as labeled by Leauthaud et al. [2007]), is that total magnitudes are estimated in a different way in each catalog and the difference is most apparent in extended objects. The magnitude estimate MAG_AUTO is an output of SExtractor that is computed with a Kron-like adaptive elliptical aperture: the ellipticity, position angle, and characteristic radius are computed for each source with the source’s moments (see SExtractor’s user manual for more information). HSC on the other hand uses composite PSF-convolved model fits to compute the CModel magnitude (see chapter 3 for a description). Figure (2.2) shows that there is a strong correlation between extendedness as measured by the HSC pipeline (see chapter 3 for description of extendedness measurement in hscPipe) and the difference between MAG_AUTO and CModel magnitude.

7These objects have extreme values for their CModel flux, so they are off the scale in the diagram. See https://hsc-jira.astro.princeton.edu/jira/browse/HSC-1113
Figure 2.2: MAG\_AUTO vs. CModel Magnitude in HSC-I in COSMOS field. MAG\_AUTO is from Leauthaud et al. [2007] and CModel Magnitude is from HSC Ultra Deep processed with hscPipe 3.10.0. The dots are color-coded by their extendedness as measured in HSC-I. The solid lines show the cuts we make to improve the quality of the matching procedure and discard outliers that may have wrong deblending in HSC or other problems.
2.2 Testing Labels

As a first test of the labels in Leauthaud et al. [2007], in figure (2.3) we plot an Ex-
tendedness vs. CModel Magnitude diagram for a subset of the COSMOS deep-coadd objects in all HSC broad bands color coded with the labels. Red dots are objects labeled as galaxies in the COSMOS HST catalog, and blue dots are labeled as stars in the COSMOS HST catalog. We expect stars to be heavily clustered around zero extendedness in this diagram, bright galaxies should have large values for extended-ness, and faint galaxies should exhibit a wide range of values including small values. The distribution of galaxies and stars are consistent with these broad expectations in all five diagrams of figure (2.3) even at quite faint magnitudes, suggesting that the labels obtained through the matching procedure in §2.1 are reasonable.

As a second test of the labels in Leauthaud et al. [2007], we look at color-color diagrams of the objects we matched to the labeled catalog. Figure (2.4) shows four riz color-color diagrams, each diagram corresponds to a different magnitude cut in HSC-I. Blue dots are objects labeled as stars, and red dots are objects labeled as galaxies. These diagrams only show galaxies with $\text{Mag}_{psf} - \text{Mag}_{cmodel} < 0.02$ in HSC-I, so they compare the distributions and relative numbers of “HSC-unresolved” galaxies and stars in the riz diagram according to the labels in Leauthaud et al. [2007]. No morphological cut is done for the objects labeled as stars.

The two top panels in figure (2.4), which correspond to the brightest magnitude cuts, show what we expect for correct labels. The putative stars lie in a tight locus, and there are only a handful of unresolved galaxies with $\text{Mag}_{cmodel} < 24$. Most of the rare unresolved bright galaxies are far from the stellar locus. The bottom-left panel

---

8 A detailed description of the extendedness measure in hscPipe can be found in chapter [3] suffic it to say here that it’s the difference between the magnitude computed with the PSF and a galaxy model fitted to the object as weight functions: $\text{Mag}_{psf} - \text{Mag}_{cmodel}$. Thus, it’s expected to be zero or close to zero for point sources, and different from zero for extended sources.

9 Since stars’ spectra are, to first approximation, determined by a single parameter—the surface temperature—they must lie in a tight 1d locus in color-color diagrams.
corresponds to $24 < \text{Mag}_{\text{model}} < 25$ and it also shows what we would expect. The putative stars cluster around the stellar locus, but it’s not as tight as in the brighter cuts because of lower S/N (for this magnitude cut S/N $\gtrsim 10$, whereas for the brighter cuts S/N $\gg 10$). More unresolved galaxies begin to appear, and these do not cluster around the stellar locus like stars do. The bottom-right diagram is not as convincing as the others because of low S/N ($\approx 10$), but one can still see by eye that the blue dots cluster around the stellar locus more than the red dots do. The diagram is also consistent with the expectation that at this latitude (COSMOS is at $b = 42^\circ$ and $l = 237^\circ$) and magnitude range, unresolved galaxies outnumber stars.

Figures (2.5) and (2.6) are the same as figure (2.4) but in the other two color-color diagrams: gri and izy respectively. The same conclusions we arrived at by examining the riz diagrams in figure (2.4) apply to the gri and izy diagrams in figures (2.5) and (2.6). These latter diagrams give us confidence not only in the labels obtained by matching to the Leauthaud et al. [2007] catalog, but also in the colors measured by HSC. By inspecting these diagrams one can see that HSC produces data of high quality (specially the riz diagrams which have the highest overall S/N).

For the remainder of the thesis, we’ll assume that the labels obtained through this matching procedure are correct. In other words, we’ll ignore classification errors in Leauthaud et al. [2007] and matching errors on our part. It should be admitted that the tests presented above are not sufficient to demonstrate that the labels are correct, or even give us a way to quantify their reliability. All we’ve really shown is that these labels are of higher quality than what we would achieve using HSC data only: recall that we are plotting galaxies that are unresolved in HSC in the color-color diagrams in figures (2.4)–(2.6). Moreover, we’ll later show that even if HSC colors are used to aid in star/galaxy separation, labels as reliable as the ones we obtained through the
matching procedure remain beyond reach\textsuperscript{10}. Since we’ve established that these labels are of higher quality than what we can achieve with HSC data only, a truth table based on these labels can be used for supervised learning with HSC measurements as inputs and the labels as outputs. This will be the focus of the next two chapters.

\textsuperscript{10}In chapter 4 we show that when colors and extendedness are used simultaneously in HSC, reliable labels can’t be predicted for objects beyond $i \approx 25$ which is below the limit $i = 25.5$ in \citep{Leauthaud2007}.
Figure 2.3: Extendedness vs. CModel Magnitude diagram for COSMOS Ultra Deep objects in all HSC broad bands. Red dots are objects labeled as galaxies in the COSMOS HST catalog, and blue dots are labeled as stars in the COSMOS HST catalog. The distribution of galaxies and stars are as expected in these diagrams, suggesting that the matching procedure works and the labels are reasonable.
Figure 2.4: Color-color diagrams in HSC-R, HSC-I, and HSC-Z for the objects in figure (2.3). Blue dots correspond to objects labeled as stars and red dots correspond to objects labeled as galaxies that have $\text{Mag}_{psf} - \text{Mag}_{cmodel} < 0.02$ in HSC-I. Each diagram corresponds to a different magnitude cut. These diagrams show that blue dots tend to cluster around the stellar locus and red dots don’t, as expected for true stars and galaxies respectively.
Figure 2.5: Same as figure (2.4) but in HSC-G, HSC-R, and HSC-I.
Figure 2.6: Same as figure (2.4) but in HSC-I, HSC-Z, and HSC-Y.
Chapter 3

Measuring Extendedness

When considering the star/galaxy separation problem, the most important question about detected sources is whether or not they are resolved (extended). Being unresolved (unextended) is a necessary (but not sufficient) condition to be a star (unless the source is blended). Most star/galaxy separation techniques used in photometric surveys’ pipelines separate stars and galaxies by separating extended from unextended sources. This worked well in the past (e.g. SDSS) because most unextended sources at the limiting magnitudes of previous surveys are stars. As we’ll see in this chapter, this is not the case for HSC despite its better seeing, so we have to be more careful when we derive labels based on whether or not an object appears to be resolved.

The existence of significant numbers of unextended galaxies at faint magnitudes is a very serious limitation for star/galaxy separation methods based on extendedness. Despite this, measuring extendedness remains the most important aspect of any solution to the problem. Extendedness cuts are the best way of doing the separation to a considerable depth \((i \lesssim 24)\), and they give a significant performance boost when combined with color-based methods at faint magnitudes (see chapter IV). Because of this, we’ve studied ways of measuring extendedness. We collect and compare various popular methods from the literature and develop a novel way of doing the measure-
ment that we argue is optimal. We focus on the extendedness measurement used in the HSC pipeline because that’s what we use in later chapters.

### 3.1 Extendedness Measurement in the HSC Pipeline

HSC inherited its star/galaxy separation from SDSS \cite{Lupton2001}. In SDSS \cite{Lupton2001} used a hard cut in a scalar measurement of extendedness as a classifier. The number used to measure extendedness is the difference between the PSF Magnitude and the CModel Magnitude: $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}}$. $\text{Mag}_{\text{psf}}$ is the magnitude computed with the PSF as a weight function\footnote{In HSC, the PSF is estimated with the PSF modeling tool PSFEx \cite{Bertin2011}. PSFEx uses nearby bright stars to sample the PSF and build a model that is dependent on position in the image.}, and $\text{Mag}_{\text{cmodel}}$ is the magnitude computed with a fitted galaxy model to the object (more details below).

For a star, the expected value of $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}}$ is zero because the best-fit model should be the PSF itself (not exactly true in practice because galaxy models always have a finite radius in the computer). For resolved galaxies on the other hand, the expected value of $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}}$ is significantly different from zero. We will show with simulations that there is a clear relation between $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}}$ and the radius of the galaxy model (see \S 3.2). For reasonable galaxy profiles, the expected value of $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}}$ grows with the radius of the galaxy model; hence the use of $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}}$ as a proxy for extendedness.

How is $\text{Mag}_{\text{cmodel}}$ estimated? The HSC pipeline, like that of SDSS, uses PSF-convolved composite models to build galaxy models. A composite model is a linear combination of a de Vaucouleurs and an exponential profile\footnote{This is referred to as the poor man’s Sérsic by Professor Lupton.}. This linear combination is used because it can approximate real galaxies with a similar accuracy to that of a full Sérsic profile, and it’s much easier to work with. The Sérsic index is a highly non-linear parameter and its estimation tends to be very unstable to noise and the initial guess. The fitting is done in two steps: in the first step the best-fit de Vaucouleurs...
and exponential profiles are obtained separately, and in the second step the best-fit linear combination of the two best-fit profiles is found. The resulting model is used as a weight function to compute the flux, and this flux is then converted to a magnitude to get $\text{Mag}_{c\text{model}}$ (we can now clarify the meaning of the term CModel, the C stands for composite). It’s important to point out that the profiles are 2-dimensional, so that besides having a radius they have an ellipticity and an angle to describe their shape and orientation. Also, we reiterate that the profiles are convolved with the estimate of the PSF model at the object’s position to evaluate the goodness of fit, and therefore the models’ shapes are a true estimate of the object’s shape. Finally, we want to mention that we know from the experience of the authors of the SDSS pipeline that CModel Magnitudes turned out to be a good estimate of total magnitudes [Lupton et al., 2001]. The algorithm that chooses the radii, ellipticities, and orientations of the de Vaucouleurs and exponential profiles is a gradient-based least-square optimizer written by Dr. James Bosch at Princeton University. To avoid arithmetic underflow, the algorithm enforces a finite and positive radius by introducing a flat prior on $\log r$ that drops smoothly to zero around $e^{-3}$ pixels. The prior on ellipticity is flat in logarithmic ellipticity, i.e. $\eta = \ln (a/b)$, and the prior on orientation angle is uniform.

Figure (2.3) shows diagrams of $\text{Mag}_{c\text{model}}$ vs $\text{Mag}_{\text{psf}} - \text{Mag}_{c\text{model}}$ for objects detected in COSMOS in HSC Ultra Deep, colored by their star/galaxy labels. In the bright end it’s clear that $\text{Mag}_{\text{psf}} - \text{Mag}_{c\text{model}}$ is a very powerful discriminant between stars and galaxies. In all five diagrams however, we see that above some value for $\text{Mag}_{c\text{model}}$ we start seeing a population of galaxies with values of $\text{Mag}_{\text{psf}} - \text{Mag}_{c\text{model}}$ that are comparable to those of stars. As we explained in chapter 2 there are two

---

3Since the de Vaucouleurs and exponential profiles are fitted separately, this can’t be thought of as a bulge-disk decomposition.

4In here by total magnitude we mean the counts we would get if we had the true galaxy profile convolved with the true PSF and used that as a weight function. The CModel calculations are done near the core of the profile where most of the signal is. In the end an aperture correction is included to account for the light in the tails.
reasons for this. One is that fainter galaxies appear smaller because they are farther away and intrinsically smaller, and the other reason is that faint galaxies that are extended have a surface brightness that is below the detection threshold. Indeed, the figure shows that near HSC’s limiting magnitude most detected galaxies look like point sources. Also, the fact that there are almost no stars beyond \( \text{Mag}_{\text{cmodel}} = 25.5 \) in HSC-I is a result of the cut made in Leauthaud et al. [2007] for point sources (left panel of figure 2.1), and see chapter 2 for details.

Figure 3.1 shows histograms of \( \text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}} \) in HSC-I for the same objects that appear in figure 2.3 in four magnitude bins. This illustrates how the distribution of galaxies overlaps more and more with that of stars for fainter bins. It’s important to note the dramatic drop in star counts relative to galaxies in the fainter magnitude bins. The galaxy histograms in the two fainter magnitude bins of figure 3.1 show a clear bimodality. It could be argued that this is because the labels in the truth catalog are wrong, and the bimodality is due to stars misclassified as galaxies. However, the putative galaxies plotted in figures 2.4–2.6 satisfy \( \text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}} < 0.02 \) in HSC-I so they are precisely the putative galaxies that lie on top of stars in the histograms. As we showed in §2.2 these putative galaxies are not consistent with a sample of stars in color-color diagrams. This indicates that the bimodality is real, and we’ll confirm this conclusion with simulations in §3.2. For now we’ll just say that the bimodality is not a consequence of a bimodality in the sizes of galaxies, but due to the details of the fitting algorithm used to compute \( \text{Mag}_{\text{cmodel}} \).

We now turn to assessing the performance of the star/galaxy classification technique used in SDSS—making a hard cut on \( \text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}} \). To evaluate performance we compute two scores for each class: completeness and purity. Completeness is defined to be the fraction of objects of a given class that have the correct predicted label (i.e. the fraction of real stars that are predicted to be stars). Purity is defined to be the fraction of objects with a given predicted label that have the correct predicted
Figure 3.1: $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}}$ in HSC-I histograms for the objects in figure (2.3) in four magnitude bins. The labels and color coding are the same as in figure (2.3).

To see the dependence of the scores on magnitude we make 50 magnitude bins between $\text{Mag}_{\text{cmodel}} = 19$ and $\text{Mag}_{\text{cmodel}} = 26$, and compute the scores for each bin. Since the $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}}$ cut the HSC pipeline uses is set by eye, we’ve done the same here. We’ve computed the scores for three different choices that appear to our eyes to achieve goals in tension with one another: high completeness for stars, a balanced separation, and high purity.
for stars. Figure (3.2) shows these three cuts in HSC-I, and figure (3.3) shows the scores for the three cuts in HSC-I. On the panel for galaxies we can see that all three choices yield high purity and completeness for galaxy samples. On the other hand the panel for stars shows low scores at faint magnitude bins. Particularly worrisome is the catastrophic collapse of purity at $\text{Mag}_{\text{comodel}} \gtrsim 24$ that effectively makes samples of stars beyond that magnitude useless. By comparing the scores of the three choices it becomes clear that giving up completeness of stars for the sake of purity results in insignificant gains in purity and significant losses in completeness. The simple reason behind this problem is that there are large numbers of unresolved galaxies at the faint magnitude bins. This is the main problem this thesis, and all recent work on star/galaxy separation, is trying to alleviate.

The results of this section confirm our prior claim that the main limitation of basing star/galaxy separation on extendedness measurements is that there are many galaxies that look like point sources at the faint end. In HSC, at magnitudes fainter than $i \sim 24$, any sample of stars selected with a cut in extendedness will have $\gtrsim 10\%$ contamination from galaxies. This renders science goals that depend on faint star counts extremely difficult or even impossible to achieve. In the next section we study HSC’s extendedness measurement in conjunction with others in simulated images. We will better understand and confirm some of the results of this section. The results in the next section will also provide evidence that the extendedness measurement HSC is currently doing is nearly optimal, so we argue that any extendedness-based technique can’t have significantly better performance than what we’ve seen here. We’ll also discuss a novel extendedness measure that we maintain is optimal under some circumstances.
Figure 3.2: \( \text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}} \) cuts in HSC-I with labeled objects. Labels are obtained from the matching procedure in chapter 2. Red dots correspond to galaxies and blue dots to stars. The horizontal lines are three extendedness cuts chosen by eye: the dashed line strives for high completeness in stars, the solid line for a balanced separation, and the dotted line for high purity in stars. Figure (3.3) shows the corresponding scores in magnitude bins.
Figure 3.3: Purity and completeness for stars and galaxies obtained from hard cuts on 
Extendedness (defined as the difference between the PSF and CModel magnitudes) in 
HSC-I. The left panel is for galaxies, and the right panel is for stars. Blue lines are for 
purity, and red lines for completeness. Dotted lines correspond to a very conservative 
cut (avoids galaxy contamination on stars), solid lines correspond to a typical cut 
chosen by eye, and dashed lines to a very permissive cut (avoids missing stars).
### 3.2 Simulating Extendedness Measurements

In this section, we’ll study the behaviour of various measures of extendedness with simulated images. The goal is to achieve a better understanding of the statistical behaviour of these measurements, and compare the various methods to discern the strengths and weaknesses of each choice. Images of both stars and galaxies are simulated with GalSim [Rowe et al., 2015], which is an open-source package for simulating images intended to study weak lensing measurements. GalSim’s source code is freely available to the public. The images produced by GalSim are then processed by the measurement module meas_multifit of hscPipe plus some additional routines that we wrote to make the measurements hscPipe doesn’t provide.

Because we want to focus on the measurements of extendedness without worrying about potential complications encountered with real data, the simulations are quite simplified. The images are prepared in such a way that only the CModel module of hscPipe and other simple computations are required for processing. Each image contains only one object in the center, so there’s no need to run a detection or deblending algorithm. The object’s centroid is treated as an input so no centroiding algorithm is used. The background’s mean and variance and the PSF’s shape are assumed to be known exactly so neither background subtraction nor PSF estimation are required. The only thing that’s unknown to the measurement routine is the profile of the object used to generate the image. Another important simplification in these simulations is that the galaxies have no substructure (no HII regions, bars, or spiral-arms), they are all generated with linear combinations of smooth de Vaucouleurs and exponential profiles.

The PSFs are chosen to be double Gaussians with parameters’ values that are typical of HSC images in the broad band filters (see figures (3.6)–(3.10)). The photometric zeropoint is chosen such that the total counts for a magnitude 0 object are

https://github.com/GalSim-developers/GalSim
The background noise is determined by matching (by eye) the distribution of objects in a $\text{Mag}_{\text{psf}}$ vs its S/N in the simulations and in real HSC data for the stars in our truth table. Figure (3.4) shows how this is done for HSC-I wide depth stacks. The pixel scale is taken to be the reported value by the Subaru Telescope for HSC, 0.17". The noise introduced by the camera’s electronics is also modeled with the values reported by the Subaru Telescope: a CCD gain of $3.0 \, \text{e}^-/\text{ADU}$, and a read noise of $4.5 \, \text{e}^-$.

Figure 3.4: Calibration of simulation’s background noise using real HSC data. The black dots show $\text{Mag}_{\text{psf}}$ vs its S/N for the stars in our truth table in HSC-I wide depth coadds. The red dots show $\text{Mag}_{\text{psf}}$ vs its S/N for simulated point sources using the background level that was deemed appropriate (by eye) to match the $\text{Mag}_{\text{psf}}$ vs S/N relation in the HSC Wide layer.

As we said in §3.1 the first extendedness measure we study is the one currently used by hscPipe: $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}}$. The second measure of extendedness is a new measure we propose that is inspired in part by the separation technique used by Leauthaud et al. [2007], MU_MAX-MAG_AUTO (see chapter 2), and $\text{Mag}_{\text{psf}} -$ 

\[\text{http://subarutelescope.org/Observing/Instruments/HSC/parameters.html}\]
Mag\textsubscript{cmodel}. Since hscPipe does not provide the output MU\_MAX, we estimate the peak surface brightness of a source with the best-fit CModel. This estimate uses all the source’s photons to estimate the peak surface brightness, while MU\_MAX uses only the brightest pixel. With the peak surface brightness at hand, an option is to follow Bertin and Arnouts [1996] and multiply by a suitable area to get a quantity analogous to MU\_MAX, instead we take advantage of the PSF model and compute the flux in the PSF model with the same peak surface brightness the CModel measured: $F_{\text{peak}}$. Finally, we call the magnitude associated with $F_{\text{peak}}$ Mag\textsubscript{peak} and use Mag\textsubscript{peak} − Mag\textsubscript{cmodel} as the extendedness measurement.

There’s an illuminating interpretation of Mag\textsubscript{psf} − Mag\textsubscript{cmodel} and Mag\textsubscript{peak} − Mag\textsubscript{cmodel} that provides us with an intuitive understanding of what each of these is doing. It’s also the basis for our claim that Mag\textsubscript{peak} − Mag\textsubscript{cmodel} is an optimal measure of extendedness. We start with a simple 2D profile and a PSF: the profile is a circular exponential with a half-light-radius of 0.5", and the PSF is a double Gaussian with FWHM = 0.8", $b = \sigma_{\text{out}}/\sigma_{\text{in}} = 2$ and $f = \text{peak}_{\text{out}}/\text{peak}_{\text{in}} = 1/10$. We convolve the profile with the PSF and depict it with the solid lines in the two top panels of figure (3.5). Now, on the top left panel we draw the best-fit PSF with a dashed line and on the top right panel we draw the PSF with the same peak surface brightness. The flux difference $F_{\text{profile}} − F_{\text{psf}}$ (the blue volume minus the red volume on the left panel) is closely related to Mag\textsubscript{psf} − Mag\textsubscript{profile}. To see how consider the following

\[
\begin{align*}
\text{Mag}_{\text{psf}} &- \text{Mag}_{\text{profile}} = \frac{-5}{2} \log_{10} \left( \frac{F_{\text{psf}}}{F_{\text{profile}}} \right) \\
&= \frac{-5}{2} \log_{10} \left( 1 - \frac{F_{\text{profile}} - F_{\text{psf}}}{F_{\text{profile}}} \right) \\
&\approx \frac{5}{2 \ln(10)} \frac{F_{\text{profile}} - F_{\text{psf}}}{F_{\text{profile}}}, \\
\end{align*}
\]
where the last step assumes $F_{\text{profile}} - F_{\text{psf}} \ll F_{\text{profile}}$ which is true for unresolved and marginally resolved sources (these are precisely the cases we are interested in). In the same way we arrive at

$$\text{Mag}_{\text{peak}} - \text{Mag}_{\text{profile}} \approx \frac{5}{2 \ln(10)} \frac{F_{\text{profile}} - F_{\text{peak}}}{F_{\text{profile}}}. \quad (3.2)$$

Viewed in this way, it’s clear that if CModel describes the profile of the source’s images accurately then $\text{Mag}_{\text{peak}} - \text{Mag}_{\text{cmodel}}$ is the most efficient way of estimating the amount of extra-flux a source has by virtue of being extended. This is what we mean by an optimal measure of extendedness, and it’s now clear that the optimal qualification is contingent on both the PSF model and the CModel fit being accurate. Mathematically, the condition to satisfy is $F_{\text{cmodel}} \approx F_{\text{profile}}$, and hereafter we assume it’s satisfied and exchange $F_{\text{profile}}$ for $F_{\text{cmodel}}$. We also point out a potential shortcoming of $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}}$ that this perspective makes apparent. If the red volume matches the blue volume we get $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}} = 0$ for an extended source. There’s even a possibility of having the red volume be larger than the blue volume and then we get $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}} < 0$ for an extended source (this explains the red dots below zero in figures (2.3) and (3.2)). The situation is of course aggravated in cases where CModel fails to capture flux in the tails, causing the estimate of the difference between the blue and the red volume to be biased low. On the other hand $\text{Mag}_{\text{peak}} - \text{Mag}_{\text{cmodel}}$ is guaranteed to be a finite positive number for extended sources, and it works even if CModel fails to capture flux in the tails of marginally resolved low S/N galaxies.

We’ll include two additional measures of extendedness. The first is a measure of extendedness called SPREAD\_MODEL. It has been implemented in SExtractor [Bertin and Arnouts, 1996] and is discussed by Bouy et al. [2013] in section 7.4 of
their paper, it’s given by

\[
\text{SPREAD\_MODEL} = \frac{G^T I}{G^T \phi} - \frac{\phi^T I}{\phi^T \phi}
\] (3.3)

where \(\phi\) is the PSF, \(G\) is the PSF convolved with a circular exponential profile with a FWHM of 1/16 of that of the PSF, \(I\) is the pixel values in the image, and \(G^T I\) denotes the inner product between \(G\) and \(I\) (i.e., summation over pixelwise products). SPREAD\_MODEL is similar to \(\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}}\) except for two things. One is that SPREAD\_MODEL uses a fixed weighting function to weight the pixel values in the first term whereas \(\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}}\) uses the best-fit CModel in the corresponding term, the second term is just the PSF flux which appears in both cases. The other difference is that SPREAD\_MODEL reports the difference in fluxes instead of the difference in magnitudes. The proponents of SPREAD\_MODEL argue that using a number that is linear in pixel values results in more desirable noise properties for the extendedness measurement, we’ll address this in our results below. In here we report the number SPREAD\_MODEL/Flux, i.e., the fractional difference, because when measuring extendedness we are not concerned with the total flux of a source but only its shape.

All the extendedness measures we’ve considered so far, use a model of the PSF. There are however star/galaxy separation techniques that don’t rely on a PSF model and are in use today. An example we have already encountered here is the method used in Leauthaud et al. [2007] (chapter 2). These techniques have the disadvantage that the threshold value depends on seeing, and so it has to be determined as a function of seeing. For instance, had the PSF of HST/ACS been different the box in the left panel of figure (2.1) would have been drawn at a different location. This is not a concern for us because the catalog of Leauthaud et al. [2007] has single-band data
that was obtained with HST/ACS, and therefore the PSF is expected to be stable.

Despite the drawbacks of not using the PSF, these techniques are still used because they are easy to implement and can work reasonably well. Although we favor the use of a PSF model if one is available, we’ll include an extendedness measure that makes no use of the PSF for a more complete comparison of the techniques used by the community.

Another example of an extendedness measure that does not use a model of the PSF can be found in MegaPipe [Gwyn, 2008]: the image processing pipeline used at the Canadian Astronomical Data Centre for MegaCam data [Boulade et al., 2003]. MegaPipe uses the half-light radius estimate from SExtractor as a discriminant for stars and galaxies. To emulate this measurement we’ll use the trace radius computed from the object’s moments \( r_{tr} = \sqrt{1/2(I_{xx} + I_{yy})} \). We use a GalSim routine that computes the adaptive moments described in Hirata and Seljak [2003] using an iterative scheme, and then use these moments to get the trace radius. We would find similar results with a half-light radius obtained by assuming a profile and multiply \( r_{tr} \) by an appropriate constant.

Since seeing will be changing in our simulations, we’ll have to locate the corresponding value of \( r_{tr} \) for point sources. Looking at the distribution of sources in \( r_{tr} \) against apparent magnitude and locating the locus of point sources, is tantamount to modeling the PSF and estimating the expected value for point sources. To make a fair comparison, we have to assume that this has been done and \( r_{tr} \) values for a point source are known. We’ll compute \( r_{tr} \) for the PSF and subtract that from the values measured for the actual sources. That is, we’ll report the quantity \( r_{tr} - (r_{tr})_{PSF} \), where

\[ 7 \text{Position-dependent charge diffusion, optical aberrations, and temperature changes along an HST orbit all cause the PSF to vary across the image and time. A Telescope Science Institute technical report on the ACS/WFC instrument estimates a variation on the order of 10\% on the FWHM of the PSF in the timescale of one orbit. Given that the median FWHM in ACS F814W is 0.12", this level of variation is far from making the HST seeing comparable to that of HSC so we don’t need to be concerned about it. See http://www.stsci.edu/hst/acs/documents/handbooks/current/c05_imaging7.html.} \]
the term with the PSF subscript is the value that we measure for the PSF. As it is the case with the other three extendedness measures we’ve considered, the expected value of $r_{tr} - (r_{tr})_{PSF}$ is zero for point sources; this makes the comparisons among the measures easier to interpret.

As a first comparison of the four extendedness measures, we’ve looked at the sensitivity of each of the extendedness measures to the parameters in the double Gaussian PSF model FWHM, and $b = \text{peak}_{out}/\text{peak}_{in}$. The parameter $f = \sigma_{out}/\sigma_{in}$ is fixed to 2, the default configuration in hscPipe. To determine a range for FWHM and $b$ that matches HSC’s PSFs, we’ve fitted double gaussians to the PSFs estimated for HSC data collected so far in the wide layer, holding $f = 2$. We allow the ellipticity and angle of the double Gaussians to vary with the constraint that they have to be the same for the inner and outer Gaussians. Figures 3.6 through 3.10 show histograms with the results of this fitting procedure, from this we get the range $0.4''$–$1.0''$ for FWHM and $0.1$–$0.3$ for $b$. We’ve taken exponential and de Vaucouleurs profiles with half-light radii between $0''$ and $0.1''$, convolved them with double Gaussian PSFs with the said range of parameters and measured the four extendedness measures discussed above: $\text{Mag}_{psf} - \text{Mag}_{cmodel}$, $\text{Mag}_{peak} - \text{Mag}_{cmodel}$, SPREAD\_MODEL, and $r_{tr} - (r_{tr})_{PSF}$. Note that no noise is added to these images, so the obtained result is what one would obtain in the limit of infinite S/N, allowing us to understand the sensitivity to changes in the PSF shape alone. Figures 3.11 and 3.12 show the extendedness measures as a function of profile radius for various choices of PSF. In figure 3.11 we vary the FWHM and keep everything else constant, and in figure 3.12 we vary $b$ and keep everything else constant.
Figure 3.5: Illustration of $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}}$ and $\text{Mag}_{\text{peak}} - \text{Mag}_{\text{cmodel}}$ extendedness measures. The source profile (solid lines) is an exponential profile with a half-light-radius of $0.5''$ convolved with a double Gaussian PSF with FWHM $= 0.8''$, $b = \sigma_{\text{out}}/\sigma_{\text{in}} = 2$, and $f = \text{peak}_{\text{out}}/\text{peak}_{\text{in}} = 1/10$. The used PSF profiles in each case are depicted by the dashed lines. The regions where the source’s profile is greater than the PSF’s are filled with blue, and the regions where the PSF’s profile is greater than the source’s are filled with red. The flux differences are equal to the blue volume minus the red volume, and the ratio between the flux difference and the source’s flux is the quantity related to the extendedness measure (See §3.2). The 2D profiles are rotated about the central pixel to generate the 3D volumes, so the tails contribute to the flux difference more than the 2D blue regions in the top panels suggest. In the bottom panels we show the profiles weighted by $2\pi r$ to get 2D areas that scale with the flux difference.
Figure 3.6: Histograms of double Gaussian parameters FWHM and $b$ in HSC-G wide. The parameters were obtained by running the hscPipe shapelet module for a double Gaussian with $f = 2$ using the same configuration the pipeline uses when approximating the PSF with shapelets, except that we force the model to be a linear combination of two zero-order shapelets, i.e., simple Gaussians.
Figure 3.7: Same as figure (3.6) but in HSC-R.
Figure 3.8: Same as figure (3.6) but in HSC-I.
Figure 3.9: Same as figure (3.6) but in HSC-Z.
Figure 3.10: Same as figure (3.6) but in HSC-Y.
Figure 3.11: The four extendedness measures as a function of half-light radius (without convolving with PSF) for circular exponential (solid lines) and de Vaucouleurs (dashed lines) profiles convolved with double Gaussian PSFs. Different colors correspond to different PSFs. No noise is added to these images. The quantity that varies for the different PSFs is the FWHM and the corresponding values are: 0.4" blue, 0.6" red, 0.8" green, and 1.0" brown.
Figure 3.12: Same as figure (3.11) but the PSFs have FWHM = 0.5” and the quantity that varies for the different PSFs is $b = \frac{\text{peak}_{\text{out}}}{\text{peak}_{\text{in}}}$. The corresponding values are: $b = 0.1$ blue, 0.17 red, 0.24 green, and 0.3 brown. All extendedness parameters are robust to changes in $b$. 
These results show very similar behaviour for all the considered extendedness measures. The extendedness measures are sensitive to the PSF’s FWHM while they’re not to $b$. This suggests that for purposes of star/galaxy separation we can consider the PSF to be a profile described by one parameter—the FWHM. The results also show that extendedness measurements of an exponential and de Vaucouleurs profiles have a different dependence on the profile radius $R_e$. Extendedness measurements for exponential profiles yield smaller values than that of de Vaucouleurs profiles for small $R_e$ but grow faster with $R_e$, yielding larger values than that of de Vaucouleurs for a large enough $R_e$.

Now we study the sensitivity of the extendedness measures to noise in the image. For this we prepare images as we did before to produce figures (3.11) and (3.12), and add noise such that S/N=30: computed using the true profile and noise model, so this number should be an upper bound on the expected S/N hscPipe gets with the best-fit CModel (a S/N=30 corresponds roughly to $i = 24$ in HSC Wide). For this we used a single PSF, a double Gaussian PSF with FWHM = 0.5", $f = \sigma_{\text{out}}/\sigma_{\text{in}} = 2$, and $b = \text{peak}_{\text{out}}/\text{peak}_{\text{in}} = 0.15$. We performed the extendedness measurements before and after adding noise. We generated 200 realizations of the noise for each profile, so that we get 200 samples of each of the extendedness measures to study their statistical behaviour. Figures (3.13) and (3.14) show our results. The dashed lines are the values computed before adding noise to the images. The solid lines are the mean values of the 200 noisy measurements at each radii, and the dotted lines are the median values. The grey regions are obtained by computing the rms variations about the mean value, for positive and negative variations separately (so it is actually a confidence interval). The distributions of these extendedness measures for a fixed profile and PSF under noise realizations can be significantly skewed. These figures show a noise bias in $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}}$ and $\text{Mag}_{\text{peak}} - \text{Mag}_{\text{cmodel}}$ that is most prominent at small $R_e$. We believe these biases are a symptom of a noise bias in the radius estimated by the fitting code.
in CModel: the radius is a positive-definite non-linear parameter and therefore it’s subject to noise biases when estimated with maximum-likelihood techniques. As we’ll see later, this noise bias in the end doesn’t affect star/galaxy separation performance.

Figure (3.15) shows $\sigma^{-1} \frac{\partial \text{ext}}{\partial R_e}$ as a function of $R_e$, where ext is the noiseless value of any of the four extendedness measures, and $\sigma$ is the width of the rms interval. This quantity is a proxy for the sensitivity of the extendedness measure to noise. Different colors correspond to different extendedness measures: blue for $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}}$, red for $\text{Mag}_{\text{peak}} - \text{Mag}_{\text{cmodel}}$, green for SPREAD\_MODEL/Flux, and brown for $r_{tr} - (r_{tr})_{PSF}$. Solid lines are for exponential profiles and dashed lines for de Vaucouleurs profiles.

---

*The main concern is that this bias in radius can introduce a bias in flux estimates; however, due to aperture corrections and calibration procedures this question becomes more subtle and less important in the case of real data.*
Figure 3.13: The four extendedness measures as a function of half-light radius (without convolving with PSF) for circular exponential profiles. The PSF used is a double Gaussian with FWHM = 0.5”, $f = \sigma_{\text{out}}/\sigma_{\text{in}} = 2$, and $b = \text{peak}_{\text{out}}/\text{peak}_{\text{in}} = 0.15$. The dashed lines are the values computed before adding noise to the images. The solid lines are the mean values for 200 realizations of the noise at each half-light radius, and the dotted lines are the median values. The noise level is chosen such that S/N=30 (using the true profile and true noise level). The grey regions are obtained by computing the rms variations about the mean value (note that these are not necessarily symmetric about the mean because of the skewness of the distributions, see text for details).
Figure 3.14: Same as figure (3.13) but using circular de Vaucouleurs profiles.
Figure 3.15: Extendedness sensitivity to noise as a function of input radius. Different colors are for different extendedness measures: blue for Mag_{psf} – Mag_{cmodel}, red for Mag_{peak} – Mag_{cmodel}, green for SPREAD_MODEL/Flux, and brown for r_{tr} – (r_{tr})_{PSF}. Solid lines are for exponential profiles and dashed lines for de Vaucouleurs profiles.
From the results in figures (3.13) through (3.15) we conclude that the four extendedness measures considered have a similar sensitivity to noise. Because of this, we expect a similar star/galaxy separation performance from any of these ways of measuring extendedness. We now further test this with a more complete comparison: we simulated a catalog of stars and galaxies and processed it with the same routines we used above to measure the simulated exponential and de Vaucouleurs profiles. To get a realistic distribution of apparent magnitudes for stars, we took all the objects labeled stars in our truth table (see chapter 2) and sampled from that list. On the other hand, to simulate galaxies we took the bulge-disc fits from the GREAT3 catalog [Mandelbaum et al., 2014], which is based on data from HST/ACS images of the COSMOS field. The bulge-disc fits of GREAT3 were obtained with the fitting method of Lackner and Gunn [2012]. In this bulge-disc decomposition the bulge is modeled as a de Vaucouleurs profile and the disk as an exponential profile. Besides radii and weights, the fitting solves for ellipticity and orientation and we include these in our simulations. The GREAT3 catalog only uses galaxies with F814W < 23.5; however, Mandelbaum et al. [2014] mention that one can use this catalog to approximate a catalog with F814W < 25.2 by rescaling the galaxies’ sizes by a factor of 0.6. We apply this rescaling factor in our calculations to extend the sample. Finally, the relative numbers of stars and galaxies are chosen to be the same we have in our truth table for F814W < 25.2.

Figure (3.16) shows all the extendedness measures as a function of magnitude for images without noise (left panels) and with noise (right panels). Blue dots are for stars and red dots are for galaxies. By looking at the left panels, we can establish that all extendedness measures work well for bright magnitudes\footnote{This is due to a combination of higher S/N and more extended galaxies’ images at brighter magnitudes.}. The right panels show that all methods should work reasonably well for $i \lesssim 24$ with the noise levels used (chosen to approximate the HSC Wide layer, see figure (3.4)), beyond that is not
very clear in this figure if there are any measures that do better. Figure (3.17) shows histograms of the extendedness measures for objects with $24 < \text{Magnitude} < 25$. Blue bars are for stars and red bars are for galaxies. This figure also does not show a strong preference for any of the four extendedness measures. In all four cases we can get pure samples of stars if we are willing to accept a low completeness. Finally, figure (3.18) shows completeness and purity as a function of magnitude for samples of stars and galaxies obtained with three different cuts in each extendedness measure. This latest figure does indicate that $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cm}}$ and $\text{Mag}_{\text{peak}} - \text{Mag}_{\text{cm}}$ do a better job than the other two. We believe this has to do with how well the PSF is modeled, and the S/N (if the background noise is too large then CModel may do worse than simple apertures or a fixed profile like in SPREADMODEL). In the case of our simulations the PSF is known exactly, so this gives an unfair advantage to the extendedness measures that use the PSF—all except $r_{tr} - (r_{tr})_{PSF}$. In the next section we’ll see that with real data the measures that use the PSF don’t do significantly better than $r_{tr} - (r_{tr})_{PSF}$.

Figure (3.17) shows that the histograms with noise for $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cm}}$ and $\text{Mag}_{\text{peak}} - \text{Mag}_{\text{cm}}$ show a bimodality that is not present in the histograms without noise. This lead us to consider the possibility that there is noise bias with hscPipe’s measurement of $\text{Mag}_{\text{cm}}$ that leads to the observed pileup in real data in figure (3.1). If this is the case, then increasing the noise background should exacerbate the pileup of galaxies near $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cm}} = 0$ and $\text{Mag}_{\text{peak}} - \text{Mag}_{\text{cm}} = 0$. We do this by rerunning this simulations with a reduced exposure time of 200 sec (exposure time of a single HSC exposure). Figure (3.19) shows the results with the reduced exposure time and it’s clear that adding noise significantly increased the pileup of extended objects near unextended ones for the measures that use $\text{Mag}_{\text{cm}}$.

Since the simulations exhibit a pileup that grows with noise, and in figure (3.1) we see the pileup worsening for fainter magnitude bins; we conclude that the most
likely explanation for the pileup is a noise bias in the Mag\textsubscript{cmodel} measurement. We’ve already seen evidence for a different noise bias in figures (3.13) and (3.14), and we had also pointed out that the radii, being a positive-definite non-linear parameters, are subject to noise bias. These simulations suggest that when the noise bias is large enough to significantly affect the star/galaxy separation performance all extendedness measures are unable to produce a sample of stars (we’ll confirm this with real data in §3.3); hence, we don’t consider this bias to be a reason to prefer the extendedness measures that don’t rely on Mag\textsubscript{cmodel}.

To better understand this bimodality, and the effects of noise on the other extendedness measures, in figure (3.20) we’ve plotted each extendedness measure (right column of figure (3.19)) against its noiseless value (left column of figure (3.19)). We can see in the plots for Mag\textsubscript{psf} − Mag\textsubscript{cmodel} and Mag\textsubscript{peak} − Mag\textsubscript{cmodel} there is a buildup of galaxies at 0 for the noisy measurements. We also see how the bias observed in figures (3.13) and (3.14) comes about: small objects tend to be scattered to large values of extendedness pulling the mean away from the true value towards larger values of extendedness. This is easy to recognize by eye for the blue dots. The extendedness measures that don’t use Mag\textsubscript{cmodel} don’t suffer from this because they let objects scatter freely towards negative values, balancing the mean against objects that are scattered to larger values\textsuperscript{10}.

Finally, we use these simulations to visualize the correlations among these extendedness measures. Figures (3.21) and (3.22) plot the extendedness measures against each other without and with noise respectively. Since Mag\textsubscript{psf} − Mag\textsubscript{cmodel} and Mag\textsubscript{peak} − Mag\textsubscript{cmodel} are tightly correlated, we omit the plots of Mag\textsubscript{peak} − Mag\textsubscript{cmodel} vs SPREAD\_MODEL and \( r_{tr} - (r_{tr})_{PSF} \) to avoid redundancy. As expected, the ex-

\textsuperscript{10}As we said in §3.1 the way the hscPipe meas\_multifit module enforces a finite and positive radius is by introducing a flat prior on log \( r \) that drops smoothly to zero around \( e^{-3} \) pixels. This is done mostly to avoid arithmetic underflow; but, as it’s typically the case in optimization problems one can never do only one thing, and an unintended consequence of this prior is to increase the skewness (and with it the bias) in the estimates for radii.
tendedness measures show strong correlations among each other; specially for small galaxies and stars. For larger sources we see no correlation or even negative correlation between SPREAD\_MODEL and the other extendedness measures. This is because SPREAD\_MODEL uses a fixed profile with a FWHM that is $1/16$ of that of the PSF so for sources that are much larger than the PSF, and therefore have a significant part of their flux beyond the PSF footprint, convolving the PSF with this small profile makes very little difference in the flux estimate. In other words, SPREAD\_MODEL is not sensitive to changes in size of objects that are much larger than the PSF. This is not the case if the CModel profile is used because it adapts itself to the size of the source, as can be confirmed by looking at the correlation between $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}}$ and $r_{tr} - (r_{tr})_{\text{PSF}}$. This is of course a minor point because for purposes of star/galaxy separation we are only concerned about sources whose size is comparable to the PSF.

The general conclusion of these simulations is that the four extendedness measures considered here result in very similar performance for star/galaxy separation. Any of these four ways of measuring extendedness is appropriate and will yield nearly optimal results (recall that we argued that $\text{Mag}_{\text{peak}} - \text{Mag}_{\text{cmodel}}$ should be optimal provided that $\text{Mag}_{\text{cmodel}} \approx \text{Mag}_{\text{total}}$ and the PSF model is accurate). To test these conclusions on real data, in the next section we compare $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}}$ and $r_{tr} - (r_{tr})_{\text{PSF}}$ with real data (hscPipe has no way of measuring $\text{Mag}_{\text{peak}}$ or SPREAD\_MODEL without modifications to the source code). The conclusions of this section will be confirmed.
Figure 3.16: Extendedness measures vs Magnitude for our simulation of stars and galaxies (See text for details). Blue dots are stars, and red dots are galaxies. The panels in the left column are for images without noise, and the panels in the right column are for images with noise. The level of noise added was such that the S/N vs Mag$_{psf}$ relation for point sources matched that of HSC-I wide data (see figure (3.4)). The horizontal black lines in the right panels indicate the three chosen cuts for the comparisons.
Figure 3.17: Histograms for extendedness measures for the objects in figure (3.16) with $24 < \text{Magnitude} < 25$. Blue bars are for stars, and red bars are for galaxies. The vertical black lines in the right panels indicate the three cuts chosen for the comparisons.
Figure 3.18: Scores for cuts in extendedness cuts using the same sample as that of figure (3.16). From top to bottom: $\Delta \text{Mag}_\text{psf} - \Delta \text{Mag}_\text{cmodel}$, $\Delta \text{Mag}_\text{peak} - \Delta \text{Mag}_\text{cmodel}$, SPREAD\_MODEL/Flux, and $r_{\text{tr}} - (r_{\text{tr}})_{\text{PSF}}$. The left column is for galaxies, and the right column for stars. The line styles correspond to the lines drawn in the right panels of figures (3.16) and (3.17), and the value is indicated in the legends.
Figure 3.19: Same as figure (3.17) but the noise matches that of single HSC-I exposures.
Figure 3.20: True vs noisy extendedness measurements in simulations. The dashed line is the locus of points where the two values are the same. The sample used to generate these is the same as that of figures (3.16) through (3.19).
Figure 3.21: Correlations among extendedness measures in simulations without noise. The sample used to generate these is the same as that of figures (3.16) through (3.19).
Figure 3.22: Same as figure (3.21) but with the noise level corresponding to the HSC Wide layer.
3.3 Extendedness Measurements on HSC Data

In this section we compare the extendedness measurements that the HSC pipeline is able to produce with HSC data. As we said in the previous section only \( \text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}} \) and \( r_{\text{tr}} - (r_{\text{tr}})_{PSF} \) can be extracted with HSC software tools without modifying or adding features to the pipeline. Since in the end we’ll classify objects into stars and galaxies in the HSC data releases, we don’t attempt to change the pipeline or reduce the data ourselves and we limit ourselves to work with the outputs hscPipe provides. We’ve had discussions with the pipeline team about adding \( \text{Mag}_{\text{peak}} \) and SPREAD\_MODEL to the list of outputs, so they may be available in future data releases.

For these tests we used our truth table for HSC-I wide and full depth (up to January 2016). Figures (3.23) and (3.24) are the same as figures (3.16) and (3.17) but with the HSC-I wide and full depth data in our truth table. As we saw in the previous section with simulations, both extendedness measures appear to have similar performance when it comes to star/galaxy separation. The histograms for the magnitude bin \( 24 < \text{Mag}_{\text{cmodel}} < 25 \) in figure (3.24) are consistent with our conclusion in the previous section that if the pileup of extended sources at \( \text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}} = 0 \) is comparable to that of stars, then other extended measures will also have comparable numbers of extended sources overlapping with stars. This, we argue, means that most of the piled-up galaxies are truly unresolved and that it’s unlikely there are significant gains in tweaking with the CModel fitting code.
Figure 3.23: Same as figure (3.16) but with HSC-I data. The labels are obtained from our truth table described in chapter 2. The black lines are three hard cuts chosen by eye, the scores for these cuts are in figures (3.25) through (3.28).
Figure 3.24: Same as figure (3.17) but with HSC-I data. The labels are obtained from our truth table described in chapter 2. The black lines are three hard cuts chosen by eye, the scores for these cuts are in figures (3.25) through (3.28).
To do a more quantitative comparison of $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}}$ and $r_{tr} - (r_{tr})_{PSF}$, we’ve chosen three hard cuts for each by eye indicated by the black lines in figures (3.23) and (3.24). The dotted line aims to yield a complete sample of stars (pure sample of galaxies), the dashed line aims to yield a pure sample of stars (complete sample of galaxies), and the solid line tries to strike a balance between the two. For each of these cuts, we compare the labels predicted by the cut to the labels in the corresponding truth table and compute the four relevant scores: purity and completeness for each class. The results are shown in figures (3.25) through (3.28) in magnitude bins.

Figure 3.25: Scores for $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}}$ cuts in HSC-I full depth data in magnitude bins. The cuts are denoted by the black lines in the top panels of figures (3.23) and (3.24). The line styles (dotted, solid, dashed) in this figure correspond to the line styles in the top panels of figures (3.23) and (3.24). Blue lines are for purity and red lines for completeness.

These results are consistent with our conclusion in the previous section that the extendedness measures studied in the simulation all have similar performance. We also see that the same problem pointed out in §3.1 for hard cuts in $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}}$ is there for hard cuts in $r_{tr} - (r_{tr})_{PSF}$: low purity for faint samples of stars. In a sense we have made no progress towards solving the problem we set out to solve here; nevertheless, we have learned something useful. We’ve showed that the four methods...
Figure 3.26: Same as figure (3.25) but with $r_{tr} - (r_{tr})_{PSF}$ cuts.

Figure 3.27: Same as figure (3.25) but with HSC-I wide data.

to measure extendedness in our simulations enjoy a performance that is very close to optimal, so efforts to improve on this front are unlikely to lead to significant improvements. In chapter 4 we’ll attempt to solve the problem from another angle, namely make use of HSC colors.
Figure 3.28: Same as figure (3.26) but with HSC-I wide data.
Chapter 4

Using HSC Colors

The limitations we found for star/galaxy separation techniques that use some kind of extendedness measurements in Chapter 3 led us to investigate the possibility of using colors as an input for the classification. The first question we’d like to answer is how much information useful to discriminate between stars and galaxies can be extracted from HSC colors alone? The first section of this Chapter provides the answer, and gives us a sense on what gains to expect from HSC colors. §4.2 puts colors together with extendedness and constructs a classifier that uses all the available information for each object. Finally §4.3 compares the classification scores obtained using our methodology to those of a widely used machine learning technique.

4.1 Using Colors Only

The tool we chose to study the information content in HSC colors is a probability density estimation algorithm that works for data with noisy realizations called Extreme Deconvolution (XD) (Bovy et al. 2011b; see Appendix B). XD infers the probability density of a random variable that we can’t observe directly: instead we observe the variable plus additive Gaussian noise with zero mean and known covariance matrix (the situation we have with color measurements). To get the probability density of
the hidden variable the noise has to be deconvolved, hence the name deconvolution.

To do the density estimate, XD uses a parametric model that consists of a linear combination of $K$ multivariate Gaussians whose parameters are chosen by maximizing the likelihood of the observed data. The number of Gaussians $K$ has to be chosen by hand, and we determine it with cross-validation (see Appendix A). XD was used to separate quasars from stars in SDSS with colors [Bovy et al., 2011a].

We use XD and our truth table to estimate the probability density of stars and galaxies in the 4 HSC colors: g-r, r-i, i-z, and z-y. This can be done because the HSC pipeline reports estimates of the errors\(^1\) in the flux measurements that we use to fill in the covariance matrix of the Gaussian noise assumed by XD (see §B.1). Once the distributions of stars and galaxies in color space have been estimated, the posterior probability of being a star can be easily obtained from Bayes’ formula

$$P(\text{Star}|\mathbf{x}, \mathbf{S}) = \frac{p(\mathbf{x}|\mathbf{S}, \text{Star})P(\text{Star})}{p(\mathbf{x}|\mathbf{S})}, \quad (4.1)$$

where $\mathbf{x}$ is the vector of color measurements, and $\mathbf{S}$ is the covariance matrix of the measurement. The likelihood is given by the convolution of the density inferred for stars with XD and a multivariate Gaussian with mean zero and covariance $\mathbf{S}$, i.e.

$$p(\mathbf{x}|\mathbf{S}, \text{Star}) = p(\mathbf{v}|\text{Star}) \ast \mathcal{N}(\mathbf{0}, \mathbf{S}), \quad (4.2)$$

where $p(\mathbf{v}|\text{Star})$ is the distribution of stars’ true colors (what we would observe for stars in the absence of noise). Note that $p(\mathbf{v}|\text{Star})$ is what XD estimates. While it’s true that the measurement error is not truly Gaussian, we do know that for high S/N it’s very close to being Gaussian [e.g., Ivezić et al., 2007]\(^2\). We’ll take the

\(^1\)We don’t study the reliability of error estimates in hscPipe here. Future technical studies by the HSC collaboration will study this and other aspects of hscPipe products.

\(^2\)This is expected because of the central limit theorem. Photon counts can be modeled with a Poisson distribution, and the theorem states that as the expected number of counts gets larger the distribution of counts approaches a Gaussian distribution.
Gaussian approximation as a reasonable one, even for low S/N. In §B.1 we describe how $S$ is determined from the errors reported by the pipeline. Finally, the term in the denominator of equation (4.1) is given by

$$p(x|S) = p(x|S, \text{Star})P(\text{Star}) + p(x|S, \text{Galaxy})P(\text{Galaxy}).$$ \hspace{1cm} (4.3)

A strength of XD is that it takes measurement errors into account. Because of this, classifications for objects from images with different depths and seeings can be predicted with a single XD fit: different observing conditions result in different reported error bars that, if properly estimated, the XD model can use to adapt itself accordingly. Consequently, we can convolve XD’s fits with the noise matrix of any source in the HSC survey as in equation (4.2), regardless of the observing conditions in the images from which the object was measured. These noisy distributions are then simply plugged into Bayes’ formula to get a posterior probability.

The distributions of galaxies and stars in color-color space will vary with magnitude because of varying redshift and metallicity distributions, so we run XD on subsamples of data in magnitude bins: $18 < \text{Mag}_{c\text{model}}^{\text{HSC-I}} < 22$, $22 < \text{Mag}_{c\text{model}}^{\text{HSC-I}} < 24$, $24 < \text{Mag}_{c\text{model}}^{\text{HSC-I}} < 25$, and $25 < \text{Mag}_{c\text{model}}^{\text{HSC-I}} < 26$. Each magnitude bin gets its own star and galaxy XD fits. We use the supervised learning framework described in Appendix A, treating the number of Gaussians $K$ as a hyperparameter. A randomly selected subset with 80% of the objects in the truth table is used to train the models, and the remaining 20% is used for testing. The resulting models use $K = 10$ for stars and galaxies in each bin.

The results of fitting the XD models to the HSC Ultra Deep COSMOS data can be illustrated as 2D projections of the ellipses describing the covariance matrices of the Gaussians in color-color diagrams, these are shown in figures (4.1) through (4.4). The top panels are for stars and the lower panels are for galaxies. One has to keep
in mind that these ellipses had the noise deconvolved. In order to see what they would look like for some S/N we would need to convolve them with an appropriate Gaussian describing the noise $\mathcal{N}(0, S)$ (we can get an intuitive sense of what this would do by recalling that the convolution of two Gaussians is itself a Gaussian with a covariance matrix that is the sum of the covariance matrices of the original two Gaussians). With these density fits at hand, the only thing we are missing in order to infer posterior probabilities of being a star is a prior. We choose the prior probability of being a star to be the fraction of the objects that are stars in each magnitude bin in our truth table. As we discuss in Appendix A.4 we have to adjust this choice if we believe there are more or less stars (relative to the COSMOS field) in some other patch of the sky where we want to use these XD models to predict star/galaxy labels.

Figure 4.1: XD fit ellipses projections into color-color diagrams in the magnitude bin $18 < \text{Mag}_{\text{c model}}^{\text{HSC-I}} < 22$. The top panels are for stars, and the bottom panels are for galaxies. Recall that these ellipses had the noise deconvolved. The lower the weight for a Gaussian the more transparent it appears in this figure.
Figure 4.2: Same as figure (4.1) but in the magnitude bin $22 < \text{Mag}_{\text{model HSC-I}} < 24$.

Figure 4.3: Same as figure (4.1) but in the magnitude bin $24 < \text{Mag}_{\text{model HSC-I}} < 25$. 
Figure 4.4: Same as figure (4.1) but in the magnitude bin $25 < \text{Mag}_{\text{model}} \text{HSC-I} < 26$. 
We can now combine the priors and the XD fits to get posterior probabilities with equation (4.1). Before doing this for real data, we want to understand what the XD models do in the absence of noise. To do this we look at \(P(\text{Star}|x, S_0)\) where \(S_0\) is a covariance matrix with 0’s in all the entries. Since \(P(\text{Star}|x, S_0)\) is a function in \(x\) (the vector of colors) and \(x\) is a 4-dimensional vector, to visualize it in color-color diagrams we have to make some sort of projection. We believe that the most informative way of doing this is to marginalize the colors that won’t appear in a given diagram in the XD probability densities, and plug in those marginal distributions in the Bayes’ formula. For example, the gri projection is obtained as follows

\[
P(\text{Star}|g-r, r-i, S_0) = \frac{p(g-r, r-i|S_0, \text{Star})P(\text{Star})}{p(g-r, r-i|S_0)}, \tag{4.4}
\]

where

\[
p(g-r, r-i|S_0, \text{Star}) = \int p(g-r, r-i, i-z, z-y|S_0, \text{Star})d(i-z)d(z-y)
\]

\[
p(g-r, r-i|S_0, \text{Galaxy}) = \int p(g-r, r-i, i-z, z-y|S_0, \text{Galaxy})d(i-z)d(z-y). \tag{4.5}
\]

The top panels of figures (4.5) through (4.8) show \(P(\text{Star}|x, S_0)\) projected on color-color diagrams as we described. These figures are very useful because they summarize what the XD classifier is doing in color-color diagrams that we are familiar with, so we can develop an intuition by examining these figures. The lower panels show a subsample of the data that generated the XD models and the priors. It has to be remembered that, unlike the projected posteriors in the top panels, the data has non-zero error bars: we include the data mostly to make sure that the posterior probability predicted by the XD models is sensible. The reason \(P(\text{Star}|x, S_0) \approx 0\) far from the stellar locus where there are very few stars and galaxies is that the Gaussians that correspond to the galaxies are broader so they dominate at the tails. This is not a
concern for us because a minute fraction of the objects appear in these regions where the posterior is unlikely to be correct. Our results are only affected by regions of color-color space where there is a non-negligible amount of probability density for either stars or galaxies.

Figure 4.5: $P(\text{Star}|\mathbf{x}, S_0)$ projected on color-color diagrams (upper panels) vs data from the truth table (lower panels) in magnitude bin $18 < \text{Mag}_{\text{modelHSC-I}} < 22$. 

73
Figure 4.6: Same as figure (4.5) but in magnitude bin $22 < \text{Mag}_{\text{model} \ HSC-I} < 24$.

Figure 4.7: Same as figure (4.5) but in magnitude bin $24 < \text{Mag}_{\text{model} \ HSC-I} < 25$. 

74
Figure 4.8: Same as figure (4.5) but in magnitude bin $25 < \text{Mag}_{\text{model}}^{\text{HSC-I}} < 26$. 
The projected posteriors in the $18 < \text{Mag}_{\text{model}}^{\text{HSC-I}} < 22$ bin show that stars in both the blue and red end of the stellar locus can be recovered with their colors alone. In the $22 < \text{Mag}_{\text{model}}^{\text{HSC-I}} < 24$ and fainter bins only stars in the red end of the stellar locus can be effectively recovered, and the discriminating information seems to be coming from $r - i$. The faintest bin shows too much noise to identify a stellar locus over the galaxies, so we’ll largely ignore it. In addition, our truth table is likely to have missing stars because of the cut in peak surface brightness applied by Leauthaud et al. [2007] (see Chapter 2). In a nutshell, our XD fits are telling us that for $i \lesssim 22$ we can separate the bluest and reddest stars from galaxies with their colors alone, and for $i \gtrsim 22$ only the reddest stars are separable. The reason the bluest stars become indistinguishable from galaxies beyond $i = 22$ is that blue star-forming galaxies begin to appear in large numbers: this was expected because the star formation rate of the universe is thought be declining since $z \approx 2$ [Lilly et al., 1996, Madau et al., 1998, Hopkins and Beacom, 2006], and K corrections for the bluest starburst galaxies at observed-frame optical wavelengths tend to be modest and weakly dependent on redshift [Kinney et al., 1996].

Figure (4.9) compares the predicted posteriors for the sources in the testing dataset (the 20% that was left out during training), with the truth table to evaluate the quality of the posteriors. The left panel plots the fraction of true stars in $P(\text{Star} | \text{Colors})$ bins to verify that the posteriors behave like a probability from the frequentist point of view (the posterior is equal to the expected fraction of stars). The right panel plots $P(\text{Star} | \text{Colors})$ against $\text{Mag}_{\text{model}}^{\text{HSC-I}}$ and colors the points according to the labels in the truth table; blue for stars, and red for galaxies. The three cuts used to compute scores as a function of magnitude are indicated in this panel by the black lines. Figure (4.10) shows the scores of the cuts denoted in the right panel of figure (4.9) using the corresponding line styles.
Figure 4.9: The left panel plots the fraction of true stars in $P(\text{Star} | \text{Colors})$ bins to verify that this quantity behaves like a probability. The right panel plots $P(\text{Star} | \text{Colors})$ against $\text{Mag}_{\text{model}}$ HSC-I and colors the points according to the labels in the truth table: blue for stars, and red for galaxies. The black lines denote the cuts we use to compute the scores in figure (4.10).

Figure 4.10: Scores in magnitude bins of $P(\text{Star} | \text{Colors})$ cuts using the cuts denoted in the right panel of figure (4.9).

These results show that, as we expected, there is useful information in colors. They also show the posterior inferred by XD behaves very much like a probability. Being able to estimate and control the contamination in a sample of stars or galaxies is very important for star-counting studies (see Chapter 5). The star scores for the $P(\text{Star} | \text{Colors}) > 0.9$ cut attest to this—the purity of the star samples remains high.
even in the faintest magnitude bins. The price we pay for this is a very low completeness in samples of stars because in the faintest magnitude bins very few objects have the colors and the S/N to identify them as stars reliably. The optimal threshold is highly dependent on the science. In MW studies where the goal is to detect substructures in the stellar halo, it’s easier to correct for missing stars than for contamination from galaxies, so a good threshold is a value close to 1 to ensure a high purity at the expense of a lower completeness. On the other hand, controlling the contamination of stars in galaxy samples is of paramount importance for WL measurements so the threshold value would be closer to zero in that case.

It’s critical to keep in mind that using colors in our star/galaxy separation will inevitably result in biased samples of both stars and galaxies. This is simply because some types of stars will be more likely to be selected than others due to their different positions in color-color diagrams. When using extendedness only, this doesn’t happen because all stars are unresolved and thus at a given S/N all stars have the same probability of being selected. If colors are used, we have to make sure that we correct the biases in our samples accordingly (see Chapter 5). To illustrate this point, figure (4.11) plots stars from our truth table in the three brightest magnitude bins, color coded by their predicted $P(\text{Star} | \text{Colors})$. The figure shows a very strong bias against blue stars in the faintest bin. Above we argued that the lack of identifiable blue stars in the fainter bins is due to the evolution of star formation in the universe: there are many more compact blue star forming galaxies in the fainter magnitude bins that have colors that match those of stars at the blue end of the stellar locus, and thus distinguishing them from stars becomes increasingly difficult.
Figure 4.11: Illustration of the biases introduced in samples of stars by $P(\text{Star}|\text{Colors})$ cuts. The stars are drawn from our truth table. Each row is for a given magnitude cut, and each column is for a given color-color diagram. Since the value of $P(\text{Star}|\text{Colors})$ assigned to a star depends on its position in the stellar locus, any sample of stars derived with a cut in $P(\text{Star}|\text{Colors})$ will result in color-dependent completeness and purity (i.e. a biased sample). As the diagrams show, the biases are more severe for the fainter magnitude bins.

4.2 Combining Colors and Extendedness

In this section we’ll attempt to use simultaneously all the information available to us to do star/galaxy separation: extendedness measurements, color measurements, and error estimates of all the measurements. Our tool of choice is XD again. In Chapter 3 we described how extendedness is measured and used for star/galaxy separation. In §4.1 we illustrated how with XD we can make use of colors and the error estimates of the color measurements. All that remains is a reliable way of estimating error bars for
the extendedness measurements, and we’ll be able to easily incorporate extendedness into the XD framework.

The HSC pipeline already provides an error bar for $r_{tr}$, $(r_{tr})_{PSF}$, $Mag_{psf}$, and $Mag_{cmodel}$; since the squared errors for the extendedness measurements are given by

$$\text{Var}[Mag_{psf} - Mag_{cmodel}] = \text{Var}[Mag_{psf}] + \text{Var}[Mag_{cmodel}] - 2\text{Cov}[Mag_{psf}, Mag_{cmodel}]$$

$$= \text{Var}[Mag_{psf}] + \text{Var}[Mag_{cmodel}] - \frac{2\text{Corr}[Mag_{psf}, Mag_{cmodel}] \sqrt{\text{Var}[Mag_{psf}] \text{Var}[Mag_{cmodel}]}}{2},$$

$$\text{Var}[r_{tr} - (r_{tr})_{PSF}] = \text{Var}[r_{tr}] + \text{Var}[(r_{tr})_{PSF}] - 2\text{Cov}[r_{tr}, (r_{tr})_{PSF}]$$

$$= \text{Var}[r_{tr}] + \text{Var}[(r_{tr})_{PSF}] - \frac{2\text{Corr}[r_{tr}, (r_{tr})_{PSF}] \sqrt{\text{Var}[r_{tr}] \text{Var}[(r_{tr})_{PSF}]}}{2},$$

we only need to estimate the correlation coefficients between $r_{tr}$ and $(r_{tr})_{PSF}$ and $Mag_{psf}$ and $Mag_{cmodel}$. We’ll only study the correlation coefficient between $Mag_{psf}$ and $Mag_{cmodel}$ with the simulations from §3.2 because we can’t estimate the correlation coefficient between $r_{tr}$ and $(r_{tr})_{PSF}$ without redesigning our simulations: in our setup the PSF is known exactly and that means $(r_{tr})_{PSF}$ is not a random variable, it’s a constant number. In any case, as we saw in §3.2 and §3.3 the choice of extendedness measure is not important. All the measures we studied convey a similar amount of information to discriminate extended from unextended objects.

For the correlation coefficient estimation we generate 100 realizations of the simulated catalog shown in figure (3.16) and then we compute the correlation coefficients with the 100 different $Mag_{psf} - Mag_{cmodel}$ noisy measurements for each object in the catalog. The left panel of figure (4.12) plots a histogram of $\text{Corr}[Mag_{psf}, Mag_{cmodel}]$ breaking it up into its star and galaxy components. From the figure it’s clear that galaxy profiles produce lower correlation coefficients than stars suggesting that the correlation coefficient is related to the extendedness of the input profile. The right
panel of figure (4.12) plots Corr $[\text{Mag}_{\text{psf}}, \text{Mag}_{\text{cmodel}}]$ against the input $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}}$ (i.e. measured without noise) to visualize that relation, which is what we should expect—the less extended an object is the more its CModel flux should behave like its PSF flux.

Figure 4.12: Correlation between $\text{Mag}_{\text{psf}}$ and $\text{Mag}_{\text{cmodel}}$ in 100 realizations of the simulated catalog shown in figure (3.16). The left panel shows a histogram of the computed correlation coefficients, for stars and galaxies separately. The right panel shows the relation between extendedness of the input profile and the correlation coefficient. The extendedness of the profile is anti-correlated with the correlation coefficient.

In our calculations below we’ll use Corr $[\text{Mag}_{\text{psf}}, \text{Mag}_{\text{cmodel}}] = 0.4$, the mode of the histogram in the left panel of figure (4.12). The right panel of figure (4.12) suggests that this will overestimate the error bar for the extendedness measurement of the smallest galaxies and the stars while underestimating the error bar for the more extended sources. Since star/galaxy separation is only affected by the small galaxies, we won’t worry about underestimating the error bar for large galaxies. On the other hand, overestimating the error bar for small galaxies and stars will result in a stronger deconvolution of the noise: the XD density estimates will be more concentrated in $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}}$ than the true distribution for small sources. When we convolve the XD fits with the Gaussian noise to get the likelihood terms $p(x|S, \text{Star})$ and
In equation (4.1) we’ll overestimate the extendedness error bar in the same way, so the introduced error will be partly corrected. We believe this simple-minded approach with all its limitations is preferable to a more sophisticated modeling of the dependence of \( \text{Corr} \left[ \text{Mag}_{\text{psf}}, \text{Mag}_{\text{cmodel}} \right] \) on galaxy shape, PSF shape, S/N and so on. The potential payoff of such a modeling exercise is small while the magnitude of the risks of modeling errors degrading performance further is unknown.

With an error bar for \( \text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}} \), we can simply add it to the vector of colors \( \mathbf{x} \) and add the corresponding entries in an enlarged error covariance matrix \( \mathbf{S} \). There is a question that remains though. Which of the five measurements of \( \text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}} \) per source should we use? There is one extendedness measurement per band, so we have to either choose one or combine them. In here we’ll simply select the extendedness measurement with the highest S/N, where we calculate S/N using the error bar estimate described above (see Appendix C for a consideration of alternatives). Finally, we assume that \( \text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}} \) is independent of the colors so there are no off-diagonal terms in \( \mathbf{S} \) associated with extendedness.

As we did in §4.1, before testing our colors+extendedness XD classifier on data we visualize the XD probability density estimates. The novelty here is extendedness, so we want to look at projections of extendedness vs a color index. In §4.1 we saw that the color that carries the most discriminatory power is \( r - i \), so we project the XD Gaussian ellipses onto \( r - i \) vs \( \text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}} \). Figure (4.13) shows this for both the galaxy and the star XD fits in all four magnitude bins. We remind the reader that these ellipses had the noise deconvolved. Figure (4.14) shows a fraction of the training data used to generate the ellipses. We learn from this figure that in the two brightest magnitude bins extendedness is the feature carrying the most discriminatory power among the inputs. In the two faintest bins on the other hand, there are significant gains in combining extendedness with colors.
Figure 4.13: XD Gaussians projected onto $r - i$ vs $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{model}}$ diagrams. Each panel corresponds to a different magnitude cut, red ellipses are for galaxies and blue ellipses for stars. The higher the weight assigned to an ellipse in the XD model the less transparent the ellipse in the figure.

The results of testing the XD classifier with the 20% of the data that was left out of the training process are in figures 4.15 and 4.16. The left panel of figure 4.15 shows that the stellar posterior can be interpreted as a probability, and the right panel plots stars and galaxies (blue and red respectively) in a posterior vs magnitude diagram. The latter plot should be compared to the right panel of figure 4.9 to notice the improvement of adding extendedness to the inputs. Figure 4.16 is the punchline of this Chapter, it shows the scores for stars (right panel) and galaxies (left panel) as a function of magnitude for three choices of posterior cuts. We see that the Colors+Extendedness XD classifier achieves high purity and completeness of both...
Figure 4.14: Data used to train the XD models in figure 4.13 but with a fraction of the data used to generate the XD density estimates. Red dots are galaxies and blue dots are for stars according to our truth table. This of course is the absolute best one can do with HSC because we are using full depth data in the COSMOS field which is an Ultra Deep field of the HSC survey. If HSC wide data is used instead the performance would of course decrease. We also want to point out that this classifier can be used with HSC wide data without making any modifications besides possibly adjusting the priors if there’s reason to believe the relative fractions of stars and galaxies are significantly different from those of the COSMOS field.

Before closing this section, we want to remind the reader of the warning we brought up at the end of §4.1: using colors in star/galaxy separation leads to biased samples of
Figure 4.15: The left panel plots the fraction of true stars in $P(\text{Star}|\text{Colors + Extendedness})$ bins to verify that this quantity behaves like a probability. The right panel plots $P(\text{Star}|\text{Colors + Extendedness})$ against $\text{Mag}_{\text{model}}$ HSC-I and colors the points according to the labels in the truth table: blue for stars, and red for galaxies. The black lines denote the cuts we use to compute the scores in figure 4.16.

Figure 4.16: Scores in magnitude bins of $P(\text{Star}|\text{Colors + Extendedness})$ cuts using the cuts denoted in the right panel of figure 4.15.

We illustrated this point in figure 4.11, and now we show the analogous result for the Colors+Extendedness classifier in figure 4.17. As expected, the introduction of an extendedness measure in the inputs led to a decrease in the bias, specially for the brighter magnitude bins. Nevertheless, the bias is still present—the faintest magnitude bin shows a strong preference for red stars. Users of this classifier should
make the appropriate corrections if needed by looking at the fraction of missed stars in the truth table as a function of colors for a given posterior cut.

Figure 4.17: Illustration of the biases introduced in samples of stars by \( P(\text{Star}|\text{Colors + Extendedness}) \) cuts. Each row is for a given magnitude cut, and each column is for a given color-color diagram. The biases become more severe for the fainter magnitude bins.

4.3 Comparison to Support Vector Machines

In this section we’ll compare our Colors+Extendedness XD classifier to a Support Vector Machine (SVM). Appendix D provides a detailed introduction to SVM. Briefly, SVM is a popular machine learning framework that is used for regression and classification problems. In classification problems, SVM aims at obtaining a smooth hypersurface in the input space that separates the two classes in the best possible way. The characteristic thing about SVM is that instead of optimizing a likelihood
function, it optimizes an error function that only takes data points that can’t be explained by the model with the current parameters (the values the parameters have at the time the error function is evaluated). In the case of classification, SVM does this by drawing a tube of some width $\epsilon$ around the separation hypersurface, and all data points that are correctly labeled by the hypersurface and lie outside the tube are ignored by the error function; only the points that are misclassified and/or lie inside the tube contribute to the error. In other words, SVM ignores data points that it can already account for with some tolerance level $\epsilon$, and focuses only on those that it can’t accommodate with the desired tolerance level. This approach leads to a number of desirable properties of the resulting SVM models. For example, sparsity: usually, SVM models can accommodate most of the data points with some reasonable tolerance $\epsilon$ so that the solution depends on the few data points that fall outside of this range. SVMs are also widely used because the kernel trick can be applied. This gives SVM a lot of flexibility to match almost any shape for a separation hypersurface (see appendix D for details). The reason we do this comparison, is that we want to make sure our classifier is up to the standards of popular machine learning techniques. If the SVM does significantly better then that means there’s room for improvement, but if the two are close we would be more confident that we are close to an optimal performance.

Why not use SVM or some other well tested machine learning technique from the outset and be done with star/galaxy separation? We prefer XD over other more popular options because we are concerned that popular machine learning techniques can’t adapt to changes in relative numbers of stars and galaxies and observing conditions such as seeing, integration time, and sky brightness (see appendix A we are in particular worried about sampling bias discussed in more detail in appendix A.3). We pointed out in §4.2 that most of the information of changes in observing conditions is encapsulated in the error bars, and since XD takes them into account we are more
confident that it will be able to adapt to data taken under different conditions. XD also allows the user to tune a prior if there is reason to believe the numbers of stars and galaxies are not sufficiently close to those of the COSMOS field and the likelihood terms are not dominant enough.

It can be argued that the error bars can be added to the list of inputs in other machine learning techniques such as SVM; however, the machine learning algorithm won’t know the difference between an input and its error bar unless error bars are somehow treated differently. This is most easily achieved with a generative classifier where the errorbars can be thought of as convolutions with Gaussian noise, which is what XD does (see appendix [B]). XD interprets the error bars as we would, and it applies a convolution that is consistent with our understanding of what an error bar is. For any machine learning technique that doesn’t make a sensible distinction between error bars and inputs, predicting a label for an object with error bars outside the training range will require outright extrapolation. Arguably, XD would also be doing extrapolation in that situation; but, even if we grant this, it’ll still be the case that XD does a more principled extrapolation that is based on our practical and theoretical understanding of error bars in photometry as opposed to some learned pattern in the data that may not generalize.

For the comparison we’ll use an SVM with a Gaussian kernel. In the case of the SVM the data is standardized for training and prediction following the guidelines of appendix [A]. The kernel bandwidth \( \sigma \) and the smoothing parameter \( C \) are treated as a hyperparameters, and like we did with XD we use 80% of the data for training and leave the remaining 20% for testing and comparisons. The inputs for the SVM are the 4 HSC colors, \( \text{Mag}_{\text{cmodel}} \) HSC-I, and the highest S/N \( \text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}} \). We

---

3A generative classifier is one that models the joint distribution of the inputs \( x \) and labels \( y \), \( p(x, y) \), and then uses Bayes rule to estimate the posterior \( p(y|x) \): XD is an example of a generative classifier. A discriminative classifier on the other hand models the posterior \( p(y|x) \) directly by learning a map from \( x \) to the posterior: SVM is an example of a discriminative classifier, with the limitation that it only gives posterior values of 0 or 1.
don’t include the covariance matrix or the error bars as inputs because, as we just explained, SVM has no way of treating a covariance matrix or error bars.

Figure 4.18 plots the scores in magnitude bins of the \( P(\text{Star}|\text{Colors + Extendedness}) = 0.5 \) (solid lines) and for the SVM (dashed lines). As before, blue lines are for purity and red lines for completeness. The scores are computed on the 20\% of the data that is left out during training. The scores for galaxies are perfect for both classifiers, and SVM does slightly better for stars. It’s not surprising that SVM yields higher scores. For one thing SVM uses all the information in the apparent magnitude measurement whereas XD splits the data in magnitude bins, forgoing some information contained in the apparent magnitude\(^4\). One also has to consider that XD provides us with more information than SVM: XD gives us a posterior probability and SVM a binary classification. We could do better than SVM with XD in star purity, for example, by increasing the posterior threshold. This kind of flexibility XD offers usually comes at a price, namely lower overall performance in exchange for a posterior output that varies smoothly between 0 and 1.

We consider that the slightly better scores seen for SVM are not enough to warrant a more exhaustive search in the parameters and hyperparameters of our XD framework to improve the scores. SVM is a very widely used technique in classification problems and it’s always common to compare new classification algorithms with it; because of this, we believe our technique is up to the standards of the machine learning literature. In chapter 5 we’ll use our XD classifier to do a pilot study of the MW using photometric parallax relations for MS stars.

\(^4\)We can’t use the apparent magnitude as an input in XD, because the distribution of both stars and galaxies in apparent magnitude is not compact enough to have it well approximated by a few Gaussians [Bovy et al., 2011a].
4.4 Changes in Seeing Conditions

In this section we study the effects of changes in seeing conditions on star/galaxy separation. This section would not have been possible without the efforts of Professor Masayuki Tanaka at The National Astronomical Observatory of Japan (NAOJ). Professor Tanaka separated the exposures in the COSMOS fields into three groups according to their seeing conditions in all four broad-band filters: best seeing, median seeing, and worst seeing. Figure 4.19 shows the seeing distribution for the three groups in HSC-I. For each of the three seeing groups, enough exposures were taken to build HSC Wide coadds in all broad-band filters, and then they were reduced in the same way as the HSC January 2016 data release. Since this was done in the COSMOS field, we could redo the matching procedure from chapter 2 and produce corresponding truth tables. Figure 4.20 shows the labeled data in each of the truth tables in HSC-I. The top panels show $\text{Mag}_{\text{cmodel}}$ vs $\text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}}$, and the bottom panels
show histograms of \( \text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}} \) in the magnitude range \( 24 < \text{Mag}_{\text{cmodel}} < 25 \). As before, blue is for stars and red for galaxies.

We use these three truth tables (best seeing, median seeing, and worst seeing) to test our XD classifier trained with the HSC Ultra Deep COSMOS catalogs. Figures (4.21) through (4.23) show the posteriors as a function of magnitude for each of the truth tables. It’s plain that worse seeing is detrimental to the classification performance, as expected. Figures (4.24) through (4.26) show scores as a function of magnitude for three cuts in the XD posterior: 0.1 (dashed), 0.5 (solid), and 0.9 (dotted). These results corroborate our hopes that XD is able to adapt to different observing conditions—seeing and depth in this case. A conservative cut in the XD posterior still leads to highly pure samples of stars for all magnitude bins, and the posterior is reasonably close to a frequentist interpretation of a probability.
Figure 4.20: Truth tables generated with the best seeing (left), median seeing (middle), and worst seeing (right) HSC-I Wide coadds in the COSMOS field. The top panels show \( \text{Mag}_{\text{cmodel}} \) vs \( \text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}} \), and the bottom panels show histograms of \( \text{Mag}_{\text{psf}} - \text{Mag}_{\text{cmodel}} \) in the magnitude range \( 24 < \text{Mag}_{\text{cmodel}} < 25 \). As before, blue is for stars and red for galaxies.

Figure 4.21: XD posteriors computed on the HSC Wide best seeing truth table. Left Panel: Comparison of XD posterior to the fraction of stars in an XD Posterior bin, the dashed line corresponds to \( P(\text{Star}|\text{Colors + Extendedness}) = \text{Star Fraction} \). Right Panel: \( P(\text{Star}|\text{Colors + Extendedness}) \) vs \( \text{Mag}_{\text{cmodel}} \) HSC-I. Blue dots are for stars, and red dots are for galaxies. The horizontal lines correspond to the cuts: 0.1 (dashed), 0.5 (solid), and 0.9 (dotted).
Figure 4.22: Same as figure (4.21) but for the HSC Wide median seeing truth table.

Figure 4.23: Same as figure (4.21) but for the HSC Wide worst seeing truth table.
Figure 4.24: Completeness (red) and purity (blue) for stars (right) and galaxies (left) in $\text{Mag}_{\text{model}}$ HSC-I bins. The line styles correspond to different cuts in $P(\text{Star}|\text{Colors + Extendedness})$, and they correspond to the cuts indicated in the right panel of figure (4.21).

Figure 4.25: Same as figure (4.24) but for the HSC Wide median seeing truth table.
Figure 4.26: Same as figure (4.24) but for the HSC Wide worst seeing truth table.
We also train a SVM in the same manner as in §4.3 using the median seeing truth table to see the effects a change in seeing can have for a traditional machine learning method. Figures (4.27) through (4.29) show the comparisons between XD and SVM for each of the truth tables. We see the same we saw in figure (4.18), namely SVM does slightly better than XD in the general classification performance. This means that XD’s performance is up to the standards of popular machine learning approaches, and as we’ve showed it adapts to different observing conditions (different exposure times, sky brightness, seeing conditions and so on).

Figure 4.27: Same as figure (4.24) but comparing SVM trained with the median seeing truth table (dashed line) to XD with a 0.5 posterior cut (solid line).

---

5The effect of differing depth and/or lower S/N for a SVM is quantified and discussed with a different dataset by Fadely et al. 2012. They find that SVM’s performance is degraded significantly when predictions are made on data with lower S/N than that of the training data.
Figure 4.28: Same as figure (4.27) but for the HSC Wide median seeing truth table.

Figure 4.29: Same as figure (4.27) but for the HSC Wide worst seeing truth table.
Chapter 5

Tomography of the Milky Way

Understanding the structure of the Milky Way (MW) is one of the oldest goals of astronomy. As early as 1785 astronomer William Herschel drew a map of the MW using his handmade telescope. Herschel already conceived of the MW as a disk-shaped structure, albeit with the Sun near its center. Over a century later, in the early 1900’s, the first evidence that the disk is supported by rotation was found by Jacobus Kapteyn. Shapley [1918] refuted the galactocentric hypothesis in 1918 based on his studies of the asymmetric distribution of globular clusters, and determined that the galactic center is in fact tens of thousands of light years away from the sun. The existence of a kinematically separate population of stars in the solar neighborhood—what we currently refer to as disk and halo—was first suggested in Jan Oort’s PhD thesis [Oort, 1926], where he argued stars should be divided into low- and high-velocity stars due to an asymmetry in stellar motion that is only present in stars above $\approx 60$ km/s. The work of Baade [1944] showed that these kinematic differences were associated with differences in the Hertzprung-Russell (HR) diagram, and also with different structural components of the galaxy: Baade concluded that the low-velocity stars (population I) correspond the solar neighborhood, whereas the fast stars (population II) correspond to globular clusters. Later on, Roman [1954] connected the
high velocities with metal deficiency by finding that high-velocity stars in the solar neighborhood show a significant excess in UV flux compared to that of their low-velocity counterparts; this also implied a connection between velocity and age. These results laid the basis from which later researchers would build theoretical accounts of our Galaxy.

The first significant attempt at a chronicle of the formation of the MW came in the work of Eggen et al. [1962]. The scenario proposed by Eggen et al. [1962] was that the high-velocity low-metallicity stars were formed during the spherical collapse phase of the gas that formed the MW, and the low-velocity high-metallicity stars formed once the gas had formed a disk. The recognition that the evolution of galaxies could be more complicated than the monolithic collapse proposed by Eggen et al. [1962] due to mergers and interactions with other galaxies [e.g., Toomre, 1977], pointed to a situation in which at least some of the stars in the halo would be relics of interactions our galaxy had with other galaxies. Searle and Zinn [1978] gave credence to this idea by showing that there is no discernible metallicity gradient in the halo, and that globular clusters show a wide range of ages while their constituent stars have no spread in metallicity. These findings are difficult to reconcile with the Eggen et al. model, so a new picture began to emerge: the stellar halo consisted mostly of the relics of galaxies that were accreted by the MW after its central regions finished collapsing.

The discovery of the Sagittarius dwarf galaxy [Ibata et al., 1994], which is currently being disrupted by the MW potential, was the confirmation that the MW is accreting smaller satellite galaxies. Later work identified groups of halo stars that are coherent in velocity implying a common origin [Majewski et al., 1996, Helmi et al., 1999]—they belonged to a dwarf galaxies that, as Sagittarius, were accreted by the MW. Surveys like SDSS and the Two Micron All Sky Survey (2MASS) marked the dawn of a new era in the study of the MW halo, and their data would reveal a halo
rich in substructure that fits naturally with the dwarf-accreting MW idea. The early galactic studies with SDSS data already showed a significant amount of substructure in the halo [Ivezić et al., 2000, Yanny et al., 2000]. With more detailed work, several substructures would later be identified in these surveys: the Monoceros stream [Newberg et al., 2002], the Sagittarius stream [Newberg et al., 2002, Majewski et al., 2003], and the Virgo overdensity [Jurić et al., 2008] among others. These datasets have also enabled measurements of the MW potential shape [e.g., Belokurov et al., 2006], and gave a way to quantify the amount of substructure and shape of the stellar halo itself [e.g., Jurić et al., 2008, Bell et al., 2008]. It’s expected that next generation surveys like HSC will yield similar results: discovery of substructures, and more detailed studies of the global properties of the stellar halo.

Our theoretical understanding of the formation of the stellar halo is by and large in line with that of Searle and Zinn [1978]. One of the testable predictions of the ΛCDM cosmological model is the manner in which galaxies like the MW form. A number of high resolution simulations of the assembly of dark matter structures of a mass comparable to that of the MW have been conducted: e.g., The Millennium Simulation [Springel et al., 2005], The Aquarius Project [Springel et al., 2008], and Via Lactea II [Diemand et al., 2008]. All these simulations show a hierarchical assembly of dark matter halos (big galaxies are formed by accreting smaller galaxies). This lends further support to the idea that the stellar halo is the relics of smaller galaxies that have been accreted by the MW. However, these simulations only take dark matter into account and completely ignore baryonic matter due to the complicated physics and computational resources baryons require; of course, in order to predict what the stellar halo of a galaxy would look like given a cosmological model baryons must be included. The classical approach has been to use semi-analytic models to post-process dark-matter-only simulations, keeping track only of dark matter particles that are accreted by the much larger central dark matter structure [Bullock and Johnston,
There are also published simulations that include baryons self-consistently and stellar halo substructures can be resolved [Abadi et al., 2006, Zolotov et al., 2009, Font et al., 2011, Tissera et al., 2013]. Today, simulations such as these are powerful tools used to understand the origin of halo substructures like rings and tidal tails, to constrain the gravitational potential of the MW, and recover the accretion history of the galaxy [e.g., Johnston et al., 2008, Zolotov et al., 2010].

The development of stellar halo simulations faced problems along the way that, arguably, have not been fully resolved. It was realized early on that ΛCDM simulations would produce a far greater number of subhalos than the number of detected MW satellite galaxies [Kauffmann et al., 1993, Klypin et al., 1999, Moore et al., 1999], and increasing the resolution of the simulations only aggravated the discrepancy. Given the enormous amount of evidence for the ΛCDM model on large scales [e.g., Riess et al., 1998, Perlmutter et al., 1999, Spergel et al., 2003, Eisenstein et al., 2005, Clowe et al., 2006, Planck Collaboration et al., 2015], and that the completeness of the MW satellite census hitherto is expected to be > 90% complete to a magnitude limit of $M_V \approx -6.5$ [Tollerud et al., 2008, Walsh et al., 2009], the likely solution is in mechanisms that suppress star-formation so that the majority of subhalos remain undetectable in stars. Models that include supernovae feedback and ultraviolet (UV) radiation from the reionization period have successfully suppressed star formation in subhalos to a degree that the simulations’ counts can be reconciled with the MW satellite demographics [e.g., Governato et al., 2007]. However, it’s been suggested that the most massive subhalos produced by ΛCDM simulations are more massive than the most massive inferred subhalos in MW satellites [Boylan-Kolchin et al., 2011, 2012]. This suggestion has been disputed by an independent analysis of the Aquarius simulation that uses a different methodology [Vera-Ciro et al., 2013], but it still remains a contentious issue: the same problem has been reported in the
Local Group Garrison-Kimmel et al., 2014 and in satellites around field galaxies Papastergis et al., 2015.

To keep making progress towards a clear resolution of the discrepancies between stellar halo simulations and what’s observed for the MW, deeper studies of the MW halo need to be carried out with the next generation photometric surveys. Discovering fainter satellites and tidal debris would allow comparisons between simulations and observations at lower mass scales. Also, finding substructures farther than ever before is very valuable to reconstruct the MW accretion history as the larger the galactocentric radius the longer the dynamical timescale to relax to a smooth distribution. HSC has the potential of making progress in all these fronts, provided clean samples of faint stars can be extracted from its catalogs. In the previous chapters we developed a method to do this, and showed that it’s reliable. We’ll now use this method to extract stars from HSC catalogs produced with the January 2016 data release, and use these stars to do a pilot study of the radial distribution of stars in the MW halo.

5.1 Photometric Parallaxes

To generate a tomography of the MW halo with stars, we need to have a way of estimating the distance to each individual star using only HSC photometry. We will do so using “photometric parallaxes”, whereby the color of a star allows us to estimate its luminosity (e.g. by assuming it’s on the MS) and this its distance via the inverse square law. In here we’ll focus on main sequence (MS) stars, but want to note that the research group led by professor Masashi Chiba at Tohoku University is using blue horizontal branch (BHB) stars in HSC to achieve the same goal. Using BHB stars has the advantage that these stars are more luminous (~ 2 mags above the MS) and the probed volume is ~ 1000 times larger; however, it has the disadvantage that BHB
stars are far less numerous than MS stars. For example, only about 1% of the stars detected in SDSS are not on the main sequence [Finlator et al., 2000]. A tomography with MS stars will have the benefit of higher star counts and will reveal substructures in more detail. In the end, the two approaches are complementary, while the one probes the large scale structure of the halo the other studies the inner region in more detail.

Photometric parallaxes for MS stars are based on an old and well-established result in astronomy: MS stars that formed from the same body of gas (and thus started with the same elemental abundances) have a well-defined mapping between color and absolute magnitude. There are several photometric parallax relations available [e.g., Jurić et al., 2008, and references therein], and in order to choose an appropriate relation we first have to establish the properties of the stars we are interested in. In here we are mostly interested in the stellar halo, so we require a photometric parallax relation that works for old (∼10 Gyr) α-enhanced ([α/Fe] ∼ 0.3) metal-poor ([Fe/H] ∼ −1.5 in the inner halo) and extremely metal-poor ([Fe/H] ∼ −2.0 in the outer halo) stars [Carollo et al., 2007, Nissen and Schuster, 2010, Ivezić et al., 2012]. There is also the possibility of contamination from thick disk stars, so we’ll also consider isochrones for populations with parameters representative of the thick disk: ∼10 Gyr, [α/Fe] ∼ 0.3, and [Fe/H] ∼ −0.5 [Wyse and Gilmore, 1995, Fuhrmann, 1998]. Finally, we’ll also include isochrones for solar abundances for comparison.

We use the publicly available theoretical isochrones of [Dotter et al., 2007] for the LSST photometric system. Figure (5.1) shows the isochrones in the four LSST colors that have an analogue in HSC. It’s immediately clear from the figure that using a photometric parallax for stars at the red end of the MS is not possible unless one has an estimate of the metallicity. Unfortunately, this is not possible without the u band because the color u − g is crucial to estimate metallicity with photometry.

1http://stellar.dartmouth.edu/models/
The tightest color-magnitude relation for the range of parameters relevant to us is in the colors $r - i$ and $i - z$ in the ranges $r - i < 0.4$ and $i - z < 0.2$ respectively, wherein for a given color value the difference in absolute magnitude between an outer-halo star and a solar neighborhood star is $\lesssim 1$ mags. We’ll focus on stars that fall in either of these ranges. Besides, these stars are the most luminous MS stars, and since we are interested in the outer halo they’re the most suitable for this study.

![Isochrones from Dotter et al. (2007) in the LSST photometric system. We only look at the colors that have an analogue in HSC.](image)

Figure 5.1: Isochrones from Dotter et al. (2007) in the LSST photometric system. We only look at the colors that have an analogue in HSC.

Figure (5.2) shows a $r - i$ vs $r$ diagram of the stars in the truth table of chapter 2. The dots are color coded by the posterior probability computed with the color+extendedness classifier of §4.2. We draw a box with a dashed line in the region where we can identify stars with a high posterior probability and we can use a photo-
metric parallax relation with the best possible accuracy. Objects that fall in this box and have a high $P(\text{Star}|\text{Colors+Extendedness})$ (i.e., $P(\text{Star}|\text{Colors+Extendedness}) > 0.9$) will be included in the tomography.

![r - i vs r diagram with posterior probabilities computed with the color+extendedness classifier from Juric et al. 2008](image)

Figure 5.2: $r - i$ vs $r$ diagram with posterior probabilities computed with the color+extendedness classifier from §4.2. The box denoted with the dashed lines contains the stars used for the tomography: this box is the region where we can safely extract stars and the photometric parallax is reliable to $\lesssim 1$ mags according to the isochrones of Dotter et al. [2007].

We use one of the photometric parallax relations in Juric et al. 2008 derived in the SDSS photometric system:

$$M_r = 4.0 + 11.86(r - i) - 10.74(r - i)^2 + 5.99(r - i)^3 - 1.20(r - i)^4. \tag{5.1}$$
The blue end of this photometric parallax relation is tied to globular cluster M13, so it’s appropriate for our purposes. To turn this into a relation in HSC colors, we numerically invert the color terms used by the HSC pipeline which were estimated by Professor Naoki Yasuda at the Kavli Institute for the Physics and Mathematics of the Universe:

\[
\begin{align*}
g_{\text{hsc}} &= g_{\text{sdss}} - 0.008 - 0.084(g - r)_{\text{sdss}} - 0.007(g - r)_{\text{sdss}}^2 \\
r_{\text{hsc}} &= r_{\text{sdss}} + 0.002 + 0.013(r - i)_{\text{sdss}} - 0.031(r - i)_{\text{sdss}}^2 \\
i_{\text{hsc}} &= i_{\text{sdss}} + 0.001 - 0.169(i - z)_{\text{sdss}} - 0.014(i - z)_{\text{sdss}}^2 \\
z_{\text{hsc}} &= z_{\text{sdss}} - 0.007 + 0.014(z - i)_{\text{sdss}} + 0.015(z - i)_{\text{sdss}}^2.
\end{align*}
\] (5.2)

Additionally, we fit straight lines to the observed stellar locus in color-color space in our truth table for the color-color region we’ll use. The resulting relations are given by

\[
\begin{align*}
g - r &= 0.158 + 1.936(r - i) \\
i - z &= -0.021 + 0.564(r - i),
\end{align*}
\] (5.3)

and they are shown by the solid lines in figures (5.3) and (5.4). Our procedure to compute the absolute magnitude of a star with measured \( r - i \) (or \( i - z \)) is as follows: evaluate equation (5.3) to obtain the corresponding coordinates for \( g - r \) and \( i - z \) in the stellar locus, numerically solve for \( (g - r)_{\text{sdss}}, (r - i)_{\text{sdss}}, \) and \( (i - z)_{\text{sdss}} \) in equation (5.2), plug \( (r - i)_{\text{sdss}} \) into equation (5.1) to obtain \( M_{r,\text{sdss}} \), and finally obtain \( M_{r,\text{hsc}} \) using equation (5.2). Figure (5.5) shows the result of applying this procedure for a given \( r - i \) or \( i - z \) in the respective allowed ranges.
Figure 5.3: Linear fit to stellar locus in $gri$ diagram for the region of color-color space we’ll use for the tomography. The left panel plots the stars detected in our truth table, and the right panel plots isodensity curves estimated with a kernel density estimate using a Gaussian kernel with a bandwidth of 0.1 mags. The solid lines show the fit, and the boxes drawn with the dashed lines indicate the region we’ll use for the tomography.

Figure 5.4: Same as figure (5.3) but in $riz$ diagram.

5.2 Tomography of the COSMOS Field

In here we’ll use the data in our truth table (i.e. stars in the COSMOS field) to test the derived photometric parallaxes, and get an idea of how far into the halo we can
Figure 5.5: Photometric parallax relations in \( r-i \) and \( i-z \) (HSC colors). This relation is based on one of the relations used by Jurić et al. [2008], see text for details.

probe. We’ll use the galactocentric distance \( d \), given by

\[
d = \sqrt{d_\odot^2 + d_\oplus^2 - 2d_\odot d_\oplus \cos b \cos l},
\]

(5.4)

where \( d_\odot \) is the distance of the sun from the galactic center, \( d_\oplus \) is the distance of the star from the earth, \( b \) is the galactic latitude of the star, and \( l \) is the galactic longitude of the star. Figure (5.6) shows various measured quantities vs the estimated galactocentric distance for each star we can use for the tomography in the COSMOS field. The figure shows that we can reach galactocentric distances of \( \sim 80 \) kpc, which means that once the HSC survey is completed it will be able to probe a volume comparable to the volume probed by Bell et al. [2008] with MS turnoff (MSTO) stars in SDSS over \( \sim 8000 \) deg\(^2\)—HSC’s area is smaller by a factor of \( \sim 8 \) but it can reach farther by a factor of \( \sim 2 \) which increases the volume-per-area by a factor of \( \sim 8 \). Since star counts drop with galactocentric distance however, the HSC MS tomography won’t contain as many stars as Bell et al. [2008].
Figure 5.6: Tomography of COSMOS field using the photometric parallax relation from §5.1 for stars with $r - i < 0.4$ and $i - z < 0.2$. Stars are selected using the classifier from §4.2. The distance in the figure is the galactocentric distance.

5.3 Tomography of HSC Wide

In this section we finally perform the tomography on the full HSC Wide footprint as of January of 2016. Figure (5.7) shows the areas imaged before the January 2016 data release in galactic coordinates. Each field has its own color: blue for XMM, green for GAMA09, red for WIDE12H, cyan for GAMA15, magenta for HectoMap, yellow for VVDS, and black for AEGIS.

Before proceeding with the tomography, we test our star selection in each of the fields by plotting objects with $P(\text{Star} | \text{Colors+Extendedness}) > 0.9$ in color-color diagrams. Figures (5.8) and (5.9) show these color-color diagrams for the XMM and GAMA15 field. As expected, the selected stars lie in the stellar locus and the
Figure 5.7: Areas imaged before the January 2016 data release in galactic coordinates. Each field has its own color: blue for XMM, green for GAMA09, red for WIDE12H, cyan for GAMA15, magenta for HectoMap, yellow for VVDS, and black for AEGIS. The solid black line is the celestial equator. The dashed black line is the plane of the Sagittarius Stream as estimated in Majewski et al. [2003]. The dashed red line is the plane of the Orphan Stream as estimated in Newberg et al. [2010]. The crosses indicate the location of the Palomar 5 globular cluster and its associated stream. The plus signs denote the position of the GD-1 Stream. The black triangle is in the Virgo over-density (Virgo constellation). The black circles are the two Magellanic Clouds. The same biases we observed in our truth table are present. This gives us confidence in the selection of stars, so we’ll proceed with it without modification to the priors or configuration. We examined the other fields in the same way.
Figure 5.8: Color-color diagrams of objects with $P(\text{Star}|\text{Colors+Extendedness}) > 0.9$ in the XMM field.
Figure 5.9: Color-color diagrams of objects with $P(\text{Star} | \text{Colors} + \text{Extendedness}) > 0.9$ in the GAMA15 field.
In here we’ll only look at star counts as a function of radius, so we’ll make bins in galactocentric radius and count the number of stars that lie in each bin according with the photometric parallax. To make sure we are comparing counts from the same stellar population across radial bins, we divide the color range $0.0 < r - i < 0.4$ into 8 bins with a width of 0.05 each and count the stars in the radial bins for each of these color bins. To correct for the color- and radius-dependent biases, we use our truth table to estimate the purity and completeness for stars chosen with $P(\text{Star} | \text{Colors+Extendedness}) > 0.9$ in each of the radius and color bins. Figure (5.10) shows this. The error bars are 95% credible intervals; they are estimated by modeling each of the scores as the probability of success in a binomial distribution $p$ (for example, for purity $p$ is the probability that an object selected as a star is truly a star), and the prior for $p$ is the Jeffreys prior for the binomial distribution (a Beta distribution with parameters $\frac{1}{2}, \frac{1}{2}$, see Brown et al. [2001]). When doing the counting in the tomography, we correct by multiplying by the estimated purity and dividing by the estimated completeness (i.e., $\text{count} \times \text{purity}/\text{completeness}$). We also propagate the errorbars in purity and completeness by sampling from the posteriors of purity and completeness, and modeling the count as a Poisson random variable with the rate equal to the observed count (prior to purity and completeness corrections).

In figure (5.11) we plot counts per square degree in $r - i$ bins in bins of galactocentric radius. We do this for each field separately, and the colors correspond to those of figure (5.1). Since we want a quantity that correlates with the number density of stars, and the bins are uniform in galactocentric radius, we divide the counts by the galactocentric radius at the center of the bin. The solid lines are smooth halo profiles $\rho_H \propto (R^2 + (Z/q_H)^2)^{-n_H/2}$ where $(R, Z)$ are the galactic cylindrical coordinates and $n_H = 2.5$, and $q_H = 0.5$ [See Jurić et al., 2008]. The vertical dashed lines show the maximum galactocentric radius at which we can detect stars for each particular field in the mid point of the color bin (see equation (5.4)).

113
Figure 5.10: Scores for stars selected with $P(\text{Star} | \text{Colors} + \text{Extendedness}) > 0.9$ in each of the radius and color bins. The error bars are 95% credible intervals; they are estimated by modeling each of the scores as the probability of success in a binomial distribution $p$ (for example, for purity $p$ is the probability that an object selected as a star is truly a star), and the prior for $p$ is the Jeffreys prior for the binomial distribution [a Beta distribution with parameters ($1/2, 1/2$), see Brown et al., 2001].

This result demonstrates that with the classifier we developed in chapter 4 we can generate a tomography of the stellar halo to a galactocentric radius of $\approx 70$ kpc with $\sim 10^5$ stars (our pilot study contains $\sim 10^4$ stars in 100 deg$^2$, and the full HSC survey will cover 1,400 deg$^2$). This will mean a significant improvement over previous studies of the stellar halo with MS stars: for example, Bell et al. [2008] mapped the halo with MSTO stars to a galactocentric radius of 30 kpc. Moreover, given all the machinery developed in this thesis it’s straightforward to carry out the tomography with MS stars. The counts show that all fields’ counts can be accommodated by the
Figure 5.11: Radial profiles of the star counts per square arcsecond in the HSC wide fields considered here. The colors correspond to that of figure (5.7). We divide the counts by the galactocentric radius at the center of the bin to get something akin to a number density. The solid lines are smooth halo profiles $\rho_H \propto (R^2 + (Z/q_H)^2)^{-n_H/2}$ where $(R, Z)$ are the galactic cylindrical coordinates and $n_H = 2.5$, and $q_H = 0.5$ [See Jurić et al., 2008]. The legends on the top right show the maximum galactocentric radius at which we can detect stars for each particular field (see equation (5.4)).

smooth stellar halo model except for two fields that show significant excess counts in some bins: XMM (blue), and GAMA15 (cyan).

Figures (5.12) through (5.18) show color-magnitude diagrams for the stars used in the stellar counts in galactocentric radius bins of figure (5.11). The dashed lines denote the color-magnitude relation for stars at a fixed distance according to the photometric parallax relation in figure (5.5). The positions of the dashed lines correspond to the edges of the bins used in figure (5.11). These figures show that the two fields
with excess counts exhibit what seems like a coherent structure at a range of radii: between 20 kpc and 40 kpc for the XMM field, and between 40 kpc and 60 kpc for the GAMA15 field. We color stars in these distance ranges in red to identify them in the equatorial coordinate diagrams below.

Figure 5.12: Color-magnitude diagram of random subset of stars from the XMM field selected for the counts in galactocentric radius bins of figure (5.11). The dashed lines denote the color-magnitude relation for stars at a fixed distance according to the photometric parallax relation in figure (5.3). The positions of the dashed lines correspond to the edges of the bins used in figure (5.11). Red dots are stars in the distance range that seems to have an excess of stars with respect to the smooth stellar halo model. In the equatorial coordinates diagrams in figures (5.19) and (5.20) these stars will preserve their red color in order to be distinguishable from stars at other photometric distances.

Figures (5.19) and (5.20) show the equatorial coordinates of a random subset of the selected stars in XMM and GAMA15 respectively. Red dots are the stars that
Figure 5.13: Same as figure (5.12) but in GAMA09 field.

appear in the bins with the excess counts, and in the range where a coherent structure is visible in the color-magnitude diagram (between 20 kpc and 40 kpc for XMM, and between 40 kpc and 60 kpc for GAMA15). Black dots are the rest of the selected stars. There is no obvious clustering of points in a substructure with the available data; however, as we’ll see below there’s reason to believe that these excess counts are due to the Sagittarius stream.
Figure 5.14: Same as figure (5.12) but in WIDE12H field.
Figure 5.15: Same as figure (5.12) but in GAMA15 field.
Figure 5.16: Same as figure (5.12) but in HectoMap field.
Figure 5.17: Same as figure (5.12) but in VVDS field.
Figure 5.18: Same as figure (5.12) but in AEGIS field.
Figure 5.19: Right ascension and declination of stars in the XMM field. Red dots are the stars that appear in the bins with the excess counts, and in the range where a coherent structure is visible in the color-magnitude diagram: between 20 kpc and 40 kpc.
Figure 5.20: Same as figure (5.19) but in the GAMA15 field, and the red dots are stars between 40 kpc and 60 kpc.
Adrian Price-Whelan pointed us to the Law and Majewski [2010] N-body simulation of the Sagittarius stream, and suggested using it to test the idea that these excess stars belong to the Sagittarius stream. The Law and Majewski [2010] N-body simulation is a dark-matter-only simulation with $10^5$ particles. The simulation starts with the particles distributed according to the Plummer model [Plummer, 1911] with a Plummer radius of $(M_{\text{Sgr}}/10^9 M_\odot)^{1/3}$ kpc, and spherically symmetric and isotropic orbits. The initial position of the Plummer sphere is determined by integrating the inferred orbit of the Sagittarius galaxy backwards in time 8 Gyr for a test particle\(^2\). The Milky Way potential is modeled as a sum of a Miyamoto & Nagai Disk [Miyamoto and Nagai, 1975], a Hernquist Spheroid, and a logarithmic halo (see Law and Majewski [2010] for details on the choice of parameters and related discussions). By comparing the velocity dispersion of the simulation particles in the tidal tails for a range of masses to the measured velocity dispersion of $8.3$ km s\(^{-1}\) [Monaco et al., 2007], they estimate a mass of $6.4 \times 10^8 M_\odot$ for the Sagittarius system (including dark matter).

Figure (5.21) shows the Law and Majewski [2010] simulation’s particles on an all-sky Mollweide projection in galactic coordinates. Every dot is one of the $10^5$ particles in their simulation. The color denotes the heliocentric distance of the particle indicated by the color bar. The convex hulls of the HSC fields shown in figure (5.11) are shown with lines using the same color convention. It’s clear from this figure that the XMM (blue) and GAMA15 (cyan) fields intersect with a large number of simulation particles. Figure (5.22) shows the simulation’s particle counts in distance bins for these two fields. According to the simulation we should expect an excess of stars in the XMM field between 10 and 40 kpc, and between 40 and 60 kpc in the GAMA15 field. This is qualitatively consistent with our data.

\(^2\)The orbit of the test particle is established by combining the position of Sagittarius measured by Majewski et al. [2003], the distance estimate of Siegel et al. [2007], the radial velocity of Ibata et al. [1997], and it’s assumed that Sagittarius is currently moving toward the Galactic plane Ibata et al. [1997, Dinescu et al. 2005].
Figure 5.21: Law and Majewski [2010] simulation’s particles on an all-sky Mollweide projection in galactic coordinates. Every dot is one of the $10^5$ particles in their simulation. The color denotes the heliocentric distance of the particle indicated by the color bar. The convex hulls of the HSC fields shown in figure (5.11) are shown with lines using the same color convention. It’s clear from this figure that the XMM (blue) and GAMA15 (cyan) fields intersect with a large number of simulation particles.

To test the idea that the excess stars are constituents of the Sagittarius stream, we make a rough estimate of the excess counts predicted by the Law and Majewski [2010] simulation in the color bins we’ve used. We begin with the simulation’s mass of the Sagittarius system: $6.4 \times 10^8 M_\odot$. This implies a mass of $6.4 \times 10^3 M_\odot$ per particle. Now, we assume a mass to light ratio of $10M_\odot/L_\odot$ to get a total luminosity of $6.4 \times 10^7 L_\odot$ for the Sagittarius system. Next, we use the Chabrier [2003] initial mass function (IMF) in conjunction with the Dotter et al. [2007] isochrones for the inner halo to get the average luminosity of a population of stars. We obtain an average luminosity of $6 \times 10^{-2} L_\odot$, and so we need $10^9$ stars to match the estimate for the luminosity of the Sagittarius system which implies $10^4$ stars per particle. Then,

$3$ As discussed in Law and Majewski [2010], this mass to light ratio is implied by the mass of the present day Sagittarius dwarf of $2.5 \times 10^8 M_\odot$ in their N-body simulation, their adopted distance of 28 kpc [Siegel et al., 2007], and the apparent magnitude of the bound core of $m_V = 3.63$ measured by Majewski et al. [2003].
Figure 5.22: Law and Majewski [2010] simulation’s particle counts in distance bins for the XMM and GAMA15 fields. According to the simulation, we should expect an excess of stars in the XMM field between 10 and 40 kpc, and between 40 and 60 kpc in the GAMA15 field. This is qualitatively consistent with our data.

we estimate the probability mass function for MS stars in the color bins used in figure (5.11) by integrating the IMF over the color bins, again using the Dotter et al. [2007] isochrones in conjunction with the Chabrier [2003] IMF. The last step gives us an estimate of the number of stars per simulation particle for each of the color bins in figure (5.11). Finally, to get a prediction for the excess star counts in each color bin for the XMM and GAMA15 fields, in each bin we count the number of simulation particles with galactocentric distance smaller than the distance detection threshold for the said fields indicated by the blue and cyan vertical dashed lines in figure (5.11). Table (5.1) shows these results together with the observed excess counts of stars in each color bin: the difference between the observed counts and the counts predicted by the smooth halo models shown in the solid lines in figure (5.11). Given the crudeness of these estimates, and the uncertainties associated with the N-body simulation, we look at this comparison with great caution, and only a dramatic difference between the observed excess star counts and the prediction based on the simulation should be considered reason enough to reject the possibility that these excess star counts are due to the Sagittarius stream. Therefore, although there
Table 5.1. Excess Star Counts from Sagittarius

<table>
<thead>
<tr>
<th>Color Bin</th>
<th>XMM Simulation $\times 10^2$ stars/deg$^2$</th>
<th>HSC Observed</th>
<th>GAMA15 Simulation $\times 10^2$ stars/deg$^2$</th>
<th>HSC Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0 &lt; r-i &lt; 0.05$</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>$0.05 &lt; r-i &lt; 0.1$</td>
<td>1</td>
<td>8</td>
<td>2</td>
<td>26</td>
</tr>
<tr>
<td>$0.1 &lt; r-i &lt; 0.15$</td>
<td>2</td>
<td>8</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>$0.15 &lt; r-i &lt; 0.2$</td>
<td>3</td>
<td>8</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>$0.2 &lt; r-i &lt; 0.25$</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$0.25 &lt; r-i &lt; 0.3$</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$0.3 &lt; r-i &lt; 0.35$</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$0.35 &lt; r-i &lt; 0.4$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

are significant disagreements between the observed counts and the predicted counts in table (5.1), we do not consider these results to be irreconcilable with the Sagittarius stream hypothesis.
Appendix A

Supervised Learning

In supervised learning the classification problem involves a dataset of labeled objects \( D = \{(x_i, y_i) \mid i = 1, \ldots, N\} \) where the \( x_i \)'s are vectors with the features of the data (e.g., colors, extendedness, and errorbars) and the \( y_i \)'s are the class labels (e.g., \( y_i \in \{\text{Star, Galaxy}\} \)). The dataset \( D \) is also known as the training dataset. The goal is to use the information in \( D \) to infer a classification rule that works well for \( D \) and, more importantly, can be generalized to other datasets for which we typically have no labels. The important thing to keep in mind is that the generalization step always involves data that had no influence in how the classification rule was inferred (i.e., no overlap with the training set). In this thesis, we often use a testing dataset that is also labeled and has no overlap with the training dataset to test classifiers trained with the training dataset—hence the name testing dataset.

Unless otherwise noted, we standardize the inputs \( x \) before running the learning algorithms, shifting and rescaling the components of \( x \) so that the \( x \)'s in the training dataset \( D \) have mean 0 and variance 1. This is a common practice and it allows a fair comparison between features that vary on different scales. One has to remember to apply the same transformation to the target dataset when predicting labels.

There are three potential problems with supervised learning:
A.1 Data Snooping

Data snooping refers to the action of using data that had an influence in the learning process to estimate the performance of the resulting classifier. An obvious example of this is to use the performance in the training dataset as an estimate of the true performance. It’s well known that this gives a biased estimate of the true performance because the learning algorithm typically chooses a model among those with the best performances in the training dataset. Thus, the training performance will be biased high as an estimator of the generalized performance.

There are more subtle examples of data snooping. For example, if both the testing and training dataset are used to standardize the data, then the training dataset is used for training, and the testing dataset to estimate the performance. In this example, the testing data had an indirect but potentially significant effect on the learning process that may bias the performance estimate. In general, the testing dataset should be left out when standardizing.

To prevent snooping problems, we completely hide the testing data when training. This ensures there will be no snooping because the testing dataset has no effect on the results of the learning algorithms.
A.2 Overfitting

Overfitting occurs when the learning algorithm fits something particular about the training dataset that does not generalize to other datasets, i.e. testing or target datasets. For example, an overly-complex model may fit particular noise features of the training data that won’t repeat in other datasets. However, overfitting is not always caused by using models that are too complex. It can also take place when using models that are simpler than the function to learn. In general, overfitting means learning patterns in the data that are not generally true and only hold for the training dataset.

There are two dangers associated with overfitting: it can lead to unreliable estimates of performance like in data snooping, and it can also lead to very poor performances out of the training data. To avoid these dangers, we always use regularization hyperparameters that reduce or increment the complexity of the model on the basis of empirical estimates of performance during training (see §A.4). We also hide a fraction of the data (testing dataset) from the learning process to estimate the performance of an algorithm; if a large difference between the training performance and the testing performance is observed this is strong evidence that overfitting has occurred.

It must be avowed that the precautions we are taking (regularization hyperparameters, and hidden testing data) only reduce the risk of overfitting, they do not remove this risk completely. For example, the training and testing dataset may share characteristics that do not generalize to other datasets, and these characteristics may have a strong influence on the result of the learning process. We are not aware of any methodology that removes the risk of overfitting completely and works in general, i.e. without making any assumptions about the underlying distribution of the features and/or the properties of the function that is being learned. Since regularization hyperparameters that are determined through cross-validation seem to be a
popular choice, we’ve chosen this methodology as our strategy to mitigate the risk of overfitting.

### A.3 Sampling Bias

Sampling bias occurs when the distribution in the training features (i.e. \( p(x) \)) is significantly different from the distribution of the target data. This can be a problem because learning algorithms try to maximize the performance for the distribution of the training data. Ideally, the probability distribution of the training dataset in the features is the same to the distribution of the target dataset. When this is the case, the Vapnik-Chervonenkis (VC) inequality applies [Vapnik, 2013]: the VC inequality provides a probabilistic bound for the out-of-training performance given an observed in-training performance. However, this bound is typically very loose (even if large amounts of data have been used for training), so there are usually no guarantees that the out-of-training performance will be close to the in-training performance for practical purposes. Nevertheless, the fact that there is no VC bound if the distributions in the training and target datasets are different shows that this difference has the potential to be a problem.

In practice, it’s not always possible to have the training and the target datasets coming from the same distributions. As [Fadely et al., 2012] point out, often in Astronomy the training dataset has objects with higher S/N than the target dataset. This is an example of sampling bias, and in the case studied by [Fadely et al., 2012] it has very adverse effects on the generalized performance of the SVM they train with high S/N objects. In the case of our XD implementation (see chapter 4) an attempt is

---

1There are cases in which this is not a problem. For example, consider the star galaxy separation problem with a PSF that is a delta function (i.e. FWHM=0). In that case, the distinction between stars and galaxies will be clear, and the distribution of the learning dataset will have no impact on the conclusion we would arrive at when learning.
made to adapt to changes in S/N, and this greatly reduces the risk of the performance degradation observed by Fadely et al. [2012] due to this sampling bias.

Besides the differences in depth (i.e., S/N), there are other sources of sampling bias that we have to consider when implementing a classification rule inferred with COSMOS data to other patches of the sky. In here we list all of the sampling bias sources we’ve identified

- Changes in relative numbers of stars and galaxies.
- Changes in seeing conditions.
- Changes in depth.
- Systematic errors in zero point fluxes.

In addition, if colors are being used, there are additional sources of sampling bias

- Changes in extinction due to variations in ISM.
- Changes in distribution of metallicities of stars.
- Changes in redshift distribution of extragalactic objects.

Now we say a few words about each of these sources of sampling bias.

To account for changes in the relative numbers of stars and galaxies, we use the priors in the XD posteriors (chapter 4). However, we’ve found that changes in the priors only have an effect in the faintest magnitude bin (25 < \text{Mag}_{\text{model,HSC-I}} < 26). In the brighter magnitude bins, the likelihood terms dominate for most objects so there is no need to worry about getting the priors right for the brighter bins. To estimate the effect of changes in seeing, we produce wide stacks in the COSMOS field with bad median and good seeing, and test our XD classifier on them (see 4.4). As we said above, changes in depth (magnitude S/N relation) are handled by the convolution step in XD. We’ll ignore systematic errors in zero point fluxes for two
reasons: one is that we expect them to be small, and the other reason is that the bulk of the information to separate stars from galaxies is in the extendedness measurements and colors, and absolute fluxes on the other hand are not that important. To account for extinction we use the extinction corrections from the HSC catalogs, which are obtained with the methodology of Schlafly and Finkbeiner [2011]. To account for changes in metallicity, we split the data into magnitude bins so that the faintest stars that are likely to be associated with the halo are more or less isolated from the disk stars (see chapter 4). Finally, we’ll ignore cosmic variance and assume that in all HSC patches the redshift distribution of extragalactic objects is sufficiently close to that of the COSMOS field.

A.4 Hyperparameters

There are two kinds of parameters we determine from the training sets, parameters and hyperparameters. Parameters are determined with an optimization algorithm (i.e. the Gradient descent algorithm to minimize an error function). Hyperparameters are determined through statistical methods that make empirical estimates of risk. The particular technique we use to determine hyperparameters is called $k$-fold cross-validation.

To illustrate this, we’ll use the XD example. In the case of XD, the hyperparameters are the number of Gaussians used for the density estimate of stars $K_{\text{star}}$ and for galaxies $K_{\text{gal}}$

$$p(x|\text{Star}) = \sum_{k=1}^{K_{\text{star}}} \alpha_{\text{star},k} \mathcal{N}(x|m_{\text{star},k}, \Sigma_{\text{star},k})$$

$$p(x|\text{Galaxy}) = \sum_{k=1}^{K_{\text{gal}}} \alpha_{\text{gal},k} \mathcal{N}(x|m_{\text{gal},k}, \Sigma_{\text{gal},k})$$
where the Gaussian means $\mathbf{m}$’s and covariances $\Sigma$’s are determined with the XD optimization algorithm for fixed $K_{\text{star}}$ and $K_{\text{gal}}$, and the $K$’s are determined with $k$-fold cross-validation. Note that the hyperparameters $K_{\text{star}}$ and $K_{\text{gal}}$ control how complex the model we fit is.

The $k$-fold cross-validation technique works as follows:

- Build a grid of potential choices of hyperparameters, and for each choice in the grid repeat these three steps several times ($\sim 100$ times with $k = 3$)

  1. Split the training set into $k$ random disjoint subsets of similar sizes.

  2. Train on $k - 1$ subsets (i.e. run the optimization algorithm to find the $\mathbf{m}$’s and $\Sigma$’s that optimize the likelihoods).

  3. Compute the classification score in the remaining subset and save the result, this is the out-of-training-score.

- Train on the whole dataset with the hyperparameters that, on average, gave the best out-of-training-score, and that’s the fit that we use in the test dataset.

Because the average out-of-training-score is used to choose hyperparameters, the risk of overfitting is reduced significantly. The out-of-training-score is a much less biased estimate of performance than the training-score. Note that $k$-fold cross-validation only uses the training set, the test set is hidden through this whole process and is used in the end to get a better estimate of the performance of the model returned by $k$-fold cross validation.

For SVM (appendix D) the hyperparameters are the smoothing parameter $C$ (see equation (D.4)), and the kernel bandwidth $\sigma$ (see equation (D.11)).
Appendix B

Extreme Deconvolution

Extreme deconvolution (XD) is an extension of a popular machine learning technique that builds generative classifiers with Gaussian mixture models (GMM) [Bovy et al., 2011b]. It has been used to target quasars for spectroscopic follow-up from the SDSS photometric survey [Bovy et al., 2011a]. There are two appealing things about this method: 1) it takes into account the estimated errors in the measurements even if they are heteroskedastic; 2) it can treat objects with missing data which is often the case in photometric surveys. Because of this, this method may work when training with data that was obtained under conditions different to those for the data we are interested in classifying (S/N, and seeing; see §A.3).

To understand how XD works, one has to understand first how GMM works. GMM uses a linear combination of $K$ Gaussian distributions to model the distribution of some vector $\mathbf{x}$

$$p(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(\mathbf{x}|\mathbf{m}_k, \Sigma_k),$$

where $\sum_k \alpha_k = 1$, and $\mathcal{N}(\mathbf{x}|\mathbf{m}_k, \Sigma_k)$ is a multivariate Gaussian distribution with mean $\mathbf{m}_k$ and covariance matrix $\Sigma_k$. The weights and parameters of the Gaussians are typically obtained by maximizing the likelihood function. The number of Gaussians $K$ can be a parameter, but it’s usually treated as a hyperparameter (see appendix A.4).
Using a GMM can be interpreted as assuming that the observed data is generated by a family of Gaussian distributions, where the probability of Gaussian $k$ generating the next point is $\alpha_k$. We’ll see that this interpretation turns out to be very useful when trying to fit GMM models to data. One of the applications of GMMs is to create generative classifiers by estimating the probability density of each class separately, and using those estimates to compute posterior probabilities of class membership.

For the case of star/galaxy separation, we would estimate the distribution of stars and galaxies separately with a mixture of Gaussians. For example, to estimate the distribution of stars, we take all the stars in the training dataset to form the star dataset $D_{\text{Star}} = \{(x_i, y_i) | y_i = \text{Star}\}$, and we fit a GMM to the density distribution of $x$ in $D_{\text{Star}}$ by maximizing the likelihood

$$p(D_{\text{Star}}|\alpha_k, m_k, \Sigma_k) = \prod_{i=1}^{N_{\text{star}}} \left( \sum_{k=1}^{K} \alpha_k \mathcal{N}(x_i|m_k, \Sigma_k) \right).$$

However, this is a very hard optimization problem. Typically, the log-likelihood function is what is fed to an optimizer to have a reasonable range of values in the objective function; but, if we look at the log of equation (B.2) we see that we can’t push the log into the terms in the sum of Gaussians. The standard techniques to obtain the maximum-likelihood don’t work, so something different must be used.

An algorithm called expectation maximization (EM) \cite{Dempster1977}, is a popular choice to maximize equation (B.2). This algorithm replaces the sum of Gaussians by a product by reinterpreting the GMM through the use of latent random variables. This is how it works: assume that instead of drawing a sample $x_i$ from the GMM in equation (B.1), first we choose one of the Gaussians and then we draw the sample $x_i$ from the chosen Gaussian. Assume also that we choose the Gaussian such that Gaussian $k$ has a probability of being chosen of $\alpha_k$—i.e, equal to its weight in the GMM. It turns out that this is equivalent to sampling from the GMM directly. To
see this, assume we’ve chosen Gaussian $k$ and write down the conditional probability for $x_i$ as $P(x_i|k) = \mathcal{N}(x_i|m_k, \Sigma_k)$; however, since $P(k) = \alpha_k$ the joint probability is $P(x_i, k) = P(x_i|k)P(k) = \alpha_k\mathcal{N}(x_i|m_k, \Sigma_k)$. Finally, we marginalize over $k$ because we are interested in $P(x_i)$ and recover the GMM in equation (B.1).

To use this insight to turn the sum in the likelihood in equation (B.2) into a product, we introduce the set of latent random variables $q_{ik} = I(x_i \sim \mathcal{N}(m_k, \Sigma_k))$, where $I(A) = 1$ if $A$ is true and 0 otherwise. Therefore, $q_{ik} = 1$ if Gaussian $k$ is chosen to generate $x_i$ and 0 otherwise. With these latent variables in hand, we can rewrite the likelihood as

$$p(D_{\text{Star}}|\alpha_k, m_k, \Sigma_k, q_{ik}) = \prod_{i=1}^{N_{\text{Star}}} \left[ \prod_{k=1}^{K} (\alpha_k\mathcal{N}(x_i|m_k, \Sigma_k))^{q_{ik}} \right].$$ (B.3)

This trick simplifies the optimization problem significantly because when we look at the log-likelihood we can now push a logarithm to the inner product operator and turn both product operators into sums. However, the latent variables $q_{ik}$ need special treatment that we’ll describe below.

The EM algorithm iterates over a two step process. In the first step the parameters of the GMM model are assumed to be known and the expectation of the log-likelihood function is computed given the dataset and the assumed parameters. The only random variables under these assumptions are the set of latent variables $q_{ik}$, so we only need to compute expectations for them. The expected value of $q_{ik}$ is simply the posterior probability that $x_i$ was generated by Gaussian $k$. In the second step, new values for the parameters are obtained by optimizing the expectation of the log-likelihood with respect to the parameters. We go back to step one with the values obtained in step two and repeat until the algorithm converges. It’s been shown that each iteration will result in an increase in the value of the likelihood in equation (B.2).
The EM algorithm does have some drawbacks. The most important one is that the result to which it converges depends on the initial guess of the GMM. That is, the EM algorithm converges to local maxima. This is not surprising; given the extreme complexity of GMMs it’s expected that there are multiple choices of parameters that yield likelihood values very close to the global maximum. However, the EM algorithm is widely used because in practice it usually converges to a sensible choice of parameters for the GMM.

The \( q_{ik} \) are called latent or hidden variables because they can be thought of as random variables that are not observed but affect the outcome of the variables we do observe. Bovy et al. [2011b] extended this idea to include missing data and true colors as latent variables. For example, if the true colors of an object are \( \mathbf{v}_i \) and we assume Gaussian noise, the observed colors \( \mathbf{x}_i \) are given by

\[
\mathbf{x}_i = \mathbf{v}_i + \mathcal{N}(0, \mathbf{S}_i),
\]

where the covariance matrix \( \mathbf{S}_i \) encodes the noise properties of the measurement. If we have missing values, we can set the variance of the missing term to a very large number to reflect our ignorance of it. In this setup, the latent variables are the indicator functions \( q_{ik} \) and the true colors \( \mathbf{v}_i \).

With this idea in mind, Bovy et al. [2011b] extended the EM algorithm to the situation described in equation (B.4). They model the distribution of the true colors \( \mathbf{v} \) with a GMM and show that the distribution of the observations \( \mathbf{x} \) is also a GMM (the distribution of \( \mathbf{x}_i \) is the convolution of the distribution of \( \mathbf{v}_i \) and \( \mathcal{N}(0, \mathbf{S}_i) \)). This result allows them to get the conditional distribution of \( \mathbf{v}_i \) given \( \mathbf{x}_i \) and \( \mathbf{S}_i \), and compute the expectation of \( \mathbf{v}_i \) under this conditional distribution. Since they can compute the expectation of all the latent variables (\( \mathbf{v}_i \) and \( q_{ik} \)), they can apply the EM algorithm. This is the same as in the simple GMM case except that there are
additional latent variables $v_i$ for which we have to compute expectations in the first step of the iteration. Bovy et al. [2011b] derive algebraic expressions to plug into each of the steps of the iteration. They also make the source code of their implementation freely available \footnote{https://github.com/jobovy/extreme-deconvolution}. We’ve used their source code for our calculations.

## B.1 Determining the Covariance Matrix

To run XD, we need to provide estimates of the colors and extendedness, denoted as $x = (g - r, r - i, i - z, z - y, \text{ext})$, and an estimate of the noise properties of the measurement, denoted as $S$. To fill the covariance matrix $S$, we assume that magnitudes in different bands are independent, and that extendedness is independent of colors. Thus, the covariance matrix is filled as follows

$$
S = \begin{pmatrix}
\sigma_g^2 + \sigma_r^2 & -\sigma_r^2 & 0 & 0 & 0 \\
-\sigma_r^2 & \sigma_r^2 + \sigma_i^2 & -\sigma_i^2 & 0 & 0 \\
0 & -\sigma_i^2 & \sigma_i^2 + \sigma_z^2 & -\sigma_z^2 & 0 \\
0 & 0 & -\sigma_z^2 & \sigma_z^2 + \sigma_y^2 & 0 \\
0 & 0 & 0 & 0 & \sigma_{\text{ext}}^2
\end{pmatrix}.
$$

(B.5)
Appendix C

Combining Multiband

Extendedness Measurements

In §4.2 we use a simple procedure to combine the extendedness measurements in different bands. We choose the measurement with the highest S/N using our definition of extendedness uncertainty—equation (4.6) with $\text{Corr} \left[ \text{Mag}_{\text{psf}}, \text{Mag}_{\text{model}} \right] = 0.4$. In here we’ll partially justify this methodology by showing that our attempts to truly combine the five extendedness measurements didn’t lead to any ponderable improvements; in fact, more sophisticated approaches often exhibited lower classification performances.

We compare the highest S/N band to two linear combinations of the five extendedness measurements. The first is the mean of the five measurements assuming they
are uncorrelated when computing the error bar:

$$\text{ext} = \frac{1}{5} \left[ (\text{Mag}_{psf} - \text{Mag}_{cmodel})_g + (\text{Mag}_{psf} - \text{Mag}_{cmodel})_r + (\text{Mag}_{psf} - \text{Mag}_{cmodel})_i + (\text{Mag}_{psf} - \text{Mag}_{cmodel})_z + (\text{Mag}_{psf} - \text{Mag}_{cmodel})_y \right]$$

$$\sigma^2_{\text{ext}} = \frac{1}{5^2} \left[ \text{Var} \left[ (\text{Mag}_{psf} - \text{Mag}_{cmodel})_g \right] + \text{Var} \left[ (\text{Mag}_{psf} - \text{Mag}_{cmodel})_r \right] + \text{Var} \left[ (\text{Mag}_{psf} - \text{Mag}_{cmodel})_i \right] + \text{Var} \left[ (\text{Mag}_{psf} - \text{Mag}_{cmodel})_z \right] + \text{Var} \left[ (\text{Mag}_{psf} - \text{Mag}_{cmodel})_y \right] \right]. \quad (C.1)$$

The second is the mean of the measurements weighted by their error bars:

$$\text{ext} = \frac{1}{5} \left[ \frac{(\text{Mag}_{psf} - \text{Mag}_{cmodel})_g}{\sqrt{\text{Var} \left[ (\text{Mag}_{psf} - \text{Mag}_{cmodel})_g \right]}} + \ldots \right]$$

$$\sigma^2_{\text{ext}} = \frac{1}{5}. \quad (C.2)$$

Figure (C.1) compares the results of using these two prescriptions to combine the extendedness measurements with the approach used in the main text. It’s clear from the figure that neither of these ways of combining bands leads to any improvement. We also pursued a way of optimally weighting each band by cross-validation; however, this always resulted in performances equal or worse to that of simply choosing the highest S/N band.
Figure C.1: Scores in magnitude bins of the cut $P(\text{Star|Colors + Extendedness}) = 0.5$ using different ways of combining the multiband measurement of extendedness: highest S/N band (solid), mean of five bands (dashed), and mean weighted by error bars (dotted).
Appendix D

Support Vector Machines

Support vector machines (SVM) can be thought of as a way of fitting a linear model. In the case of classification, SVM fits a linear model of the form \( w^T x \), where \( w \) is simply a vector of weights for each data feature and \( w^T x = 0 \) is the decision boundary that separates the two classes. The classification function is defined as

\[
    f_w(x) = \begin{cases} 
    \text{Star} & \text{if } w^T x > 0 \\
    \text{Galaxy} & \text{if } w^T x \leq 0 
    \end{cases}.
\]

(D.1)

The thing we want to optimize when fitting the model is the classification error rate

\[
    \text{Err}(w) = \sum_{i=1}^{N} I(y_i \neq f_w(x_i)),
\]

(D.2)

where \( I(A) = 1 \) if \( A \) is true and 0 otherwise.

The optimization problem in equation (D.2) is basically finding the hyperplane that separates the two classes in the best way, this is a hard problem. There is an algorithm called perceptron [Rosenblatt, 1957] that tries to solve it directly with the iterative scheme

\[
    w(t + 1) = w(t) + \alpha(y_i - f_{w(t)}(x_i))x_i,
\]

(D.3)
where \(0 < \alpha < 1\) is the learning rate (set by the user), \(\mathbf{w}(t)\) is the weights vector at iteration \(t\), \((\mathbf{x}_i, y_i)\) is an element of \(\mathcal{D}\) picked randomly, and \(y_i\) and \(f_{\mathbf{w}(t)}(\mathbf{x}_i)\) take the value 1 when they give Star and 0 when they give Galaxy. The iterations continue until all the objects are perfectly separated or until some other criterion is satisfied. It’s not uncommon for the perceptron to get stuck or converge to completely different results for slightly different initial conditions. One of the things that makes this optimization problem hard is that there are typically many solutions. The trick behind SVM is to add an additional constraint on the problem to turn it into one with a unique solution.

If we consider data that can be separated perfectly by a hyperplane, we can easily see how there will be multiple hyperplanes that achieve the separation. How do we choose among them? A reasonable choice is the hyperplane that lies furthest from the data, that is the hyperplane that maximizes the distance to the closest data point. This is called the maximum margin hyperplane and it’s unique. It turns out that the problem of finding the maximum margin hyperplane can be written as a quadratic programming problem. This is very convenient because quadratic programming problems are a very well studied case of convex programming problems and many software packages have been written to solve problems of this kind.

For data that can’t be perfectly separated SVM allows for some points to be misclassified. A set of variables \(\zeta_i\) called slack variables are introduced (one for every training point). These variables take the value of 0 if the object is correctly classified and farther than some margin from the decision boundary, and a value proportional to the distance by which the margin is violated otherwise. If \(0 < \zeta_i < 1\) the object is correctly classified but its distance from the decision boundary is smaller than the specified margin, and if \(\zeta_i > 1\) the object is misclassified. The algorithm minimizes

\[
\mathbf{w}^T \mathbf{w} + C \sum_{i=1}^{n} \zeta_i, \tag{D.4}
\]
where $C$ is a hyperparameter. Since $\zeta_i > 1$ means that object $i$ is misclassified we can think of $\sum_{i=1}^{n} \zeta_i$ as an upper bound on the number of misclassified points. It turns out that this problem too has a unique solution and can be written as a quadratic programming problem.

It’s important to note that there is a trade-off between the number of misclassified objects, and the minimum distance from the decision boundary objects with $\zeta_i = 0$ have (i.e. the margin that determines when $\zeta_i > 0$). The hyperparameter $C$ controls the relative importance between these two things. A smaller $C$ means we are willing to tolerate more classification errors in order to have a decision boundary that lies further from the objects that have $\zeta_i = 0$ so that we are more confident of the predicted labels for these objects.

A convenient property about SVM is that its solutions tend to be sparse. That is, the solutions for the optimal $w$ are of the form

$$w = \sum_{i}^{N} \alpha_i x_i, \quad (D.5)$$

where $\alpha_i \geq 0$ and the inequality holds for only a few points. The few points $x_i$ for which $\alpha_i > 0$ are called support vectors, hence the name. Support vectors are the points closest to the boundary for perfectly separable data, and the points with the largest $\zeta_i$ for non-separable data. The reason the sparsity is convenient, is that we can efficiently predict labels for any point $x$ by evaluating the sum

$$f(x) = \sum_{i}^{N} \alpha_i x_i^T x, \quad (D.6)$$

and since the solution is sparse the evaluation of this sum takes very little time.

A limitation of this algorithm is that it can’t capture nonlinear relations that may be present in the classification rule, because it fits a linear model. One simple way to fix this is to define a nonlinear transformation of the data $z = \phi(x)$ and then fit the
linear model in $\mathcal{Z}$ space. The $\mathcal{Z}$ space may have more dimensions than the $\mathcal{X}$ space where the data lives. This procedure can add a lot of complexity to the model, so it must be used very judiciously. One can easily get a perfect fit by using a complex enough transformation, but this is likely to lead to overfitting.

An example of a nonlinear transformation of the data is one that uses all the first and second order terms

$$
\phi(x) = \phi((x_1, x_2, \cdots, x_D)) = (1, x_1, \cdots, x_D, x_1^2, x_1x_2, \cdots, x_ix_j, \cdots x_D^2). \quad (D.7)
$$

For example if $\mathcal{X}$ has two features $x = (x_1, x_2)$ then $\phi(x) = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2)$. If we use the five features in the star/galaxy problem ($g - r$, $r - i$, $i - z$, $z - y$, and extendedness), under this transformation $\mathcal{Z}$ has 21 features (a much more complex model!). Such a complicated model is likely to lead to overfitting unless there are huge amounts of data for training.

A better way of transforming the data to pick up nonlinear features is to use the kernel trick. In this context, a kernel is a two-argument function $\kappa(x,x') \in \mathbb{R}$ that is symmetric $\kappa(x,x') = \kappa(x',x)$, satisfies $\kappa(x,x) \geq 0$ for any $x$, and for any set of inputs $\{x_i | i = 1 \cdots N\}$ the matrix defined by

$$
K = \begin{pmatrix}
\kappa(x_1, x_1) & \cdots & \kappa(x_1, x_N) \\
\vdots & \ddots & \vdots \\
\kappa(x_N, x_1) & \cdots & \kappa(x_N, x_N)
\end{pmatrix} \quad (D.8)
$$

is positive definite. Mercer’s theorem states that for a kernel satisfying these conditions, there exists a transformation $\phi(x)$ such that

$$
\kappa(x,x') = \phi(x)^T \phi(x'). \quad (D.9)
$$
This fact is useful because in SVM, the dual of the optimization problem in equation (D.4) can be written in terms of products of the form $\phi(x)^T \phi(x')$. Also, equation (D.6) under a transformation $\phi(\cdot)$ associated with a kernel $\kappa(\cdot, \cdot)$ can be rewritten as

$$f(x) = \sum_i^N \alpha_i \phi(x_i)^T \phi(x) = \sum_i^N \alpha_i \kappa(x_i, x).$$  \hspace{1cm} (D.10)

This means that we can call the kernel function $\kappa(\cdot, \cdot)$ instead of evaluating the inner products. This is known as the kernel trick or kernelization, and it increases efficiency in cases where the transformation $\phi(\cdot)$ increases the number of features to a large number (possibly infinite!).

The kernel trick also allows us to use transformations $\phi(x)$ that take the data to a space $Z$ of infinite dimension and fit the model in finite time. This is done by using a kernel that can be written as an inner product of vectors with infinite dimension (i.e. an infinite series). A well known example of this is the radial basis function (RBF) kernel

$$\kappa(x, x') = \exp \left( -\frac{||x - x'||^2}{2\sigma^2} \right),$$  \hspace{1cm} (D.11)

which for $\sigma = 1$ can be written as

$$\exp \left( -\frac{||x - x'||^2}{2} \right) = \sum_{j=0}^{\infty} \frac{(x^T x')^j}{j!} \exp \left( -\frac{||x||^2}{2} \right) \exp \left( -\frac{||x'||^2}{2} \right).$$  \hspace{1cm} (D.12)

By examining this series, we could derive the form of the transformation $\phi(x)$ that generates this kernel. It turns out that for equation (D.12) the associated transformation takes the $x$'s to all the possible orders for multivariate Hermite polynomials weighted by a Gaussian with mean 0 and $\sigma = 1$. What’s remarkable about this, is that even though we are fitting a model in a space with an infinite number of dimensions overfitting is not as overwhelming as one would naively expect.
If we supply a kernel to a kernelized SVM, the algorithm has everything it needs
to fit the model. All we have to do then is choose a good kernel. The most important
characteristic about the kernel we use is the bandwidth ($\sigma$ in the case of the RBF
kernel), the shape is not as important. We treat kernel parameters as hyperparameters
and determine them with $k$-fold cross-validation.

To fit linear models, we use the SVM implementations in the Python module
Scikit-learn [Pedregosa et al., 2011]. The particular SVM implementation we use is
based on the C library LIBLINEAR [Fan et al., 2008] developed by the department of
computer science of the National Taiwan University. To fit kernelized models, we use
the Scikit-learn [Pedregosa et al., 2011] implementation of a kernelized SVM based
on the C library LIBSVM [Chang and Lin, 2011]. This library implements the RBF
kernel (equation (D.11)), polynomial kernels of the form $\kappa(x, x') = (\gamma x^T x' + r)^d$, and
the sigmoid kernel $\kappa(x, x') = \tanh(\gamma x^T x' + r)$. These kernels allow us to use SVM as
if we had used complicated transformations on the data without having to actually
do so. Using the polynomial kernel with $d = 2$ for example, is equivalent to applying
the transformation in equation (D.7) and then fitting a linear SVM in $\mathcal{Z}$. 
Bibliography


151


