University of Rochester

Senior Thesis

Identifying Type Ia Supernovae in Extragalactic Spectra

Author:

Ryan Rubenzahl

Supervisor:

Professor Segev BenZvi

A thesis submitted to the Department of Physics and Astronomy gaining a full enrichment of the bachelor's degree. Supervised by Professor Segev BenZvi

in the

Department of Physics and Astronomy

University of Rochester

Rochester, New York

2018

Abstract

Identifying Type Ia Supernovae in Extragalactic Spectra

by

Ryan Rubenzahl

Bachelor of Science

Department of Physics and Astronomy

Professor Segev BenZvi, Advisor

With future astronomical surveys expecting to see millions of transient events per night (such as LSST), there is a need to develop efficient means of identifying interesting objects. One such class of interesting objects are type Ia supernovae. I investigate the optical spectral footprint of type Ia supernovae with the goal of understanding how these objects may be identified within the spectra of their host galaxies. I describe and compare several machine learning and data-driven approaches for identifying type Ia supernovae spectroscopically. Finally, I detail the application of these methods to the Dark Energy Spectroscopic Instrument (DESI), which will observe 30 million galaxy spectra over the next ten years and as such is a prime instrument for finding supernovae and other interesting phenomena spectroscopically.

Contents

List of Figures

List of Tables

Acknowledgements

What a four years it has been. I would not be where I am today without the support and encouragement of the many incredible and talented people I have met and had the pleasure of working with here at Rochester. I'll try to keep it brief, since this is only an undergraduate Senior Thesis, but I still have to do this section justice.

First, I am endlessly indebted to my research advisor of four years, Professor Segev BenZvi. There is no-one I have learned more from about what it means to be a scientist and what it means to be a mentor to others, and I could not have asked for a better mentor myself. Thank you.

Additional thanks to those who helped in part with this project over the last year. Professor Douglass and Tolga, thank you for many constructive conversations and for sharing your expertise during our group meetings. To the rest of the group, Chang, Mehr, and Tri, and special thanks to Ouail, who helped with much of the preliminary work to get this project going.

An unending thanks and a tremendous "hurrah!" to my fellow PAS majors, with whom I have endured countless all-nighters, long problem sets, and overnight observing runs to Mees: Bo, Pooya, Josh, Genevieve, Lexie, and Jean. And to those who have already graduated: Tanveer, Sanha, Greg, and Chelsea. Thanks for being great teachers and friends.

There are many professors and faculty who have inspired and motivated me to continue to pursue this field. Professors McFarland, Blackman, Wolfs, Orr, Rajeev and Frank. A very special thanks is owed to Professor Dan Watson and to Dr. Eric Mamajek, without whom I would not be going to graduate school. And lastly to Janet Fogg, who confirmed my intent to come here to Rochester, although I didn't know it at the time. Thank you all.

Finally, to my family, who have provided endless support: Mom, Dad, and Evan. And to the first person I met on campus, without whom I'm not sure I would have been able to manage, thank you Katherine. I can't wait to see where the future takes us.

Chapter 1

Type Ia Supernovae

Stars do not live forever. Throughout their lifetimes stars undergo nuclear fusion reactions in their cores, producing photons and heat that drive an outward pressure capable of supporting the massive outer layers of the star under the star's own weight. Stars begin by fusing hydrogen into helium until eventually[1] the star's reserve of hydrogen in its core is exhausted. At this point, the star contracts until its central temperature and pressure are raised enough to initiate helium fusion, and the resulting energy production halts the contraction. The helium in the star's core fuses into carbon and then oxygen via the triple alpha process, and a thin layer of hydrogen surrounding the core continues fusing into helium. The star's luminosity increases and gradually pushes the surrounding envelope outwards.

The core continues to contract until the central densities become large enough that the hot gas of ionized electrons in the stellar core becomes degenerate. The degenerate electrons exert their own pressure, *electron degeneracy pressure*, which can be sufficient to halt the collapse of the star. As the star exhausts the remaining helium and hydrogen shells surrounding the core, its outer envelope continues to expand and diffuse away, resulting in a planetary nebula. The thin helium and hydrogen shells gradually extinguish, and all that remains is the degenerate core of carbon and oxygen. This object, the final stage of evolution for stars up to about 8 $M_\odot$, is called a **white dwarf**. More massive stars can fuse some of the carbon, but not the resulting neon, yielding an oxygen-neon-magnesium white dwarf. It is also possible for lower mass stars in binary systems to

[1] For a 1 $M_\odot$ star like our Sun, this stage lasts about 10 billion years.

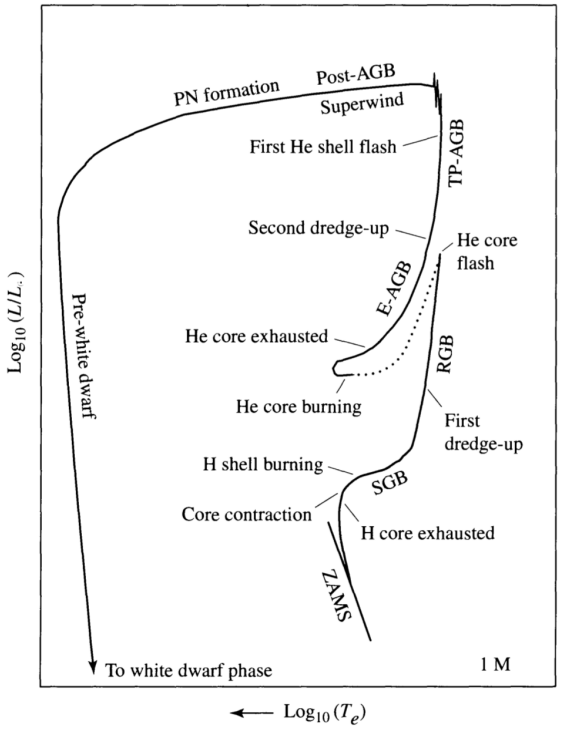

FIGURE 1.1: A diagram of the evolutionary path of a 1 $M_\odot$ over its lifetime. The star begins at the zero-age main sequence (ZAMS), eventually exhausting its hydrogen and moving onto the subgiant branch (SGB), red giant branch (RGB), early asymptotic giant branch (E-AGB), thermal pulse asymptotic giant branch (TP-AGB), and post-asymptotic giant branch (Post-AGB), leading to planetary nebula formation (PN formation) and the creation of a white dwarf. From Carroll and Ostlie, 2007.

strip the outer envelope at a faster rate, before the core reaches carbon, leaving behind a helium white dwarf. Figure 1.1 shows the evolutionary path of a 1 $M_\odot$ star over its lifetime.

1.1 Phenomenology

A white dwarf (WD) maintains hydrostatic equilibrium due to electron degeneracy pressure, a consequence of the Pauli exclusion principle. An important property of electron degeneracy pressure is that it is temperature independent and only depends on the density of electrons. If a WD is massive enough, its self-gravity can overwhelm electron degeneracy pressure and collapse the WD. The mass limit at which this occurs is known as the Stoner-Anderson-Chandrasekhar mass, or Chandrasekhar mass, and is about 1.44 $M_\odot$ for

FIGURE 1.2: Type Ia supernova SN 2011fe in the Pinwheel galaxy (M101). Credit: B. J. Fulton, Las Cumbres Observatory Global Telescope Network.

a carbon-oxygen (C-O) WD (Anderson, 1929; Stoner, 1930; Chandrasekhar, 1931).

A more massive star ($> 8\ M_\odot$) whose core exceeds $1.44\ M_\odot$ will collapse until a *neutron degenerate* core forms, in which case the in-falling material rebounds off the stiff neutron core and explodes the star outwards. This scenario is called a *core-collapse supernova* and produces the observed *Type II*, *Ib*, and *Ic* classes of supernovae. These observed classes are distinguished by their optical spectra, which will be discussed in more detail in Section 1.2. Core-collapse supernovae constitute about 75% of all observed supernovae (roughly 50% type II and 25% type Ib/c), with the remaining 25% being type Ia (Graur et al., 2017). If the star were even more massive, the neutron-degenerate core can be overwhelmed by gravity and collapse into a *black hole*.

If the C-O WD is able to increase its mass through some means, for example by accreting mass from a companion star, it can approach or even exceed the Chandrasekhar mass. As it does, the increase in mass raises the pressure and temperature of the WD until carbon fusion ignites. Since electron degeneracy pressure is independent of temperature,

the resulting heat produced by the fusing carbon does not raise the pressure and cause the WD to expand and cool, as a normal star would. Instead, the heat is trapped and increases the temperature further, which in turn increases the rate of fusion. This feedback loop creates a runaway thermonuclear reaction that releases enough energy in a short period of time to exceed the gravitational binding energy of the WD - that is, it explodes violently in a **type Ia supernova** (SN Ia).

1.1.1 Explosion Mechanism and Progenitor Models

The scenario that gives rise to a SN Ia involves some form of multi-star system, in which one of the members is a WD, typically a C-O WD. The C-O WD accretes matter from its companion until it reaches the Chandrasekhar mass and explodes, producing enough luminosity to outshine the rest of the host galaxy. The companion may be another WD (*double-degenerate scenario*), an evolved stellar companion (*single-degenerate*), or even a triple system consisting of two C-O WDs and a normal star. While the progenitors of SNe Ia have yet to be directly observed, owing to the faintness of WDs relative to their bright stellar companions, SNe Ia have been observed that give evidence both for and against each of these progenitor scenarios. Here, I will briefly outline the progenitor models in more detail and provide observational examples of each (see Branch et al., 1995 and Maoz and Mannucci, 2012 for reviews on the state of the type Ia progenitor issue).

Single Degenerate Binary System

A stellar binary system consists of two stars, a higher mass (primary) and lower mass (secondary) companion. The primary evolves off the main sequence first, proceeds through the giant phases and becomes a WD. The resulting system is now a WD - star binary system. The secondary later evolves off main-sequence, and as it enters the giant phases the WD companion can accrete matter from the secondary via Roche lobe overflow or through stellar winds. The steady flow of material from the secondary to the WD gradually raises the mass of the WD towards the Chandrasekhar mass.

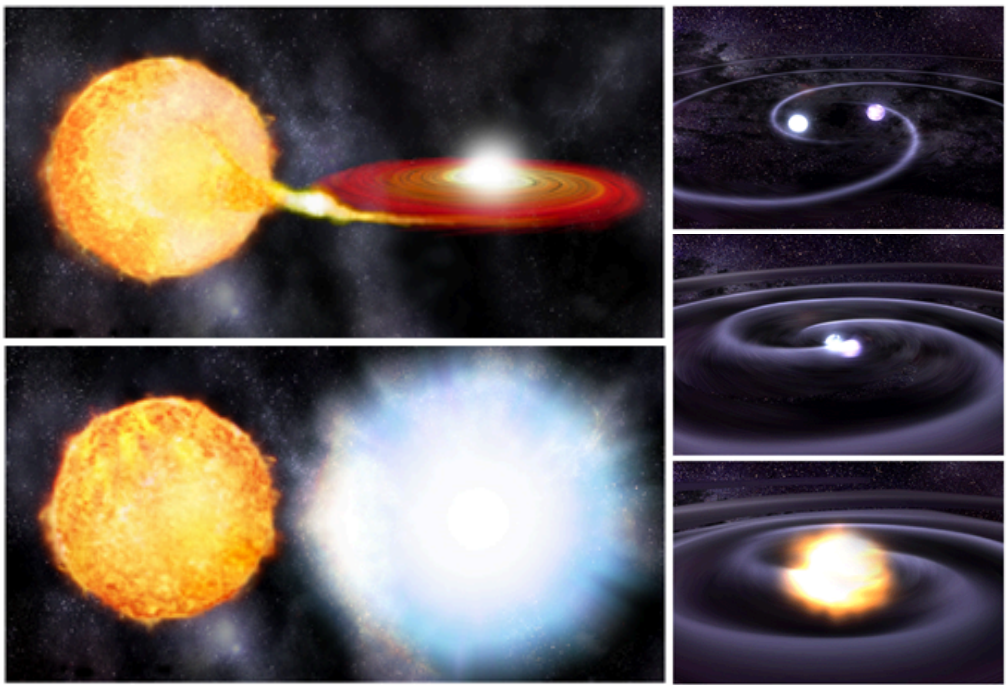

FIGURE 1.3: An artist's concept of the single degenerate (left) and double degenerate (right) progenitor models for type Ia supernovae. Credit: NASA / CXC / M. Weiss.

The actual moment in which the WD explodes as a SN Ia is actually believed to occur very near to, but not quite at, the Chandrasekhar mass. The prevailing model states that the temperature in the WD's core reaches a critical temperature at which a carbon-fusing deflagration flame ignites (Timmes and Woosley, 1992). The exact details surrounding the ignition and propagation of this flame are an active area of research (see Hillebrandt and Niemeyer, 2000 for a review of current SN Ia explosion mechanisms), but the consensus is that the flame instigates the runaway thermonuclear reaction that releases sufficient energy to unbind the WD, involving temperatures on the order of 10^9 K and densities on the order of 10^7 g cm^{-3} (Woosley, Kerstein, and Aspden, 2011).

As a majority of stars exist in binary systems, this WD-companion configuration is believed to be relatively common in the universe. For example, the type Ia supernova SN 2002ic, which exhibited Hα features in its spectrum (type Ia spectra do not contain hydrogen by definition, as we will discuss later) and non-type Ia components in its lightcurve. These anomalous features were consistent however with the presence of an asymptotic giant branch star. This evidence suggests a non-degenerate companion, supporting the

single-degenerate model.

Double Degenerate Binary System

If a stellar binary system is unable to share mass between companions, both stars may evolve into a WD without detonating the primary, yielding a WD-WD binary system. Such a system could arise by a WD orbiting within a common envelope of another star, ejecting the outer layers to reveal its companion's WD core, and placing the WDs at a relatively close orbital separation (Wang and Han, 2012). The pair of WDs will emit gravitational waves as they orbit each other, resulting in the loss of orbital energy and the decay of the orbital separation. Eventually the two WDs will inspiral and merge, producing a super-Chandrasekhar mass WD that immediately collapses and explodes.

There are a number of observational examples that support such a scenario. One of the strongest is the anomalous supernova SN 2003fg (SNLS-03D3bb). Observations of the lightcurve of SN 2003fg revealed that the progenitor contained around 1.3 $M_\odot$ of ^{56}Ni, and observations of the spectrum concluded that 40% of the material in the supernova were elements other than ^{56}Ni, implying a WD mass of $\sim$ 2.1 $M_\odot$ (Howell, Sullivan, and Nugent, 2006). Hence, the progenitor WD was well-above the Chandrasekhar mass, implying a possible merger of two (sub-Chandrasekhar) WDs.

Other examples include the type Ia supernova remnant SNR 0509-67, which has no sign of any ex-companion stars in or around the remnant nebula (Schaefer and Pagnotta, 2012). Likewise the remnant of type Ia SN 1006 shows no signs of having a surviving evolved companion (González Hernández et al., 2012). These events imply a single WD exploding alone in space, which is possible in the double-degenerate model.

However, the WD-WD binary merger rate in galactic disks does not appear to be enough to reproduce the observed SNe Ia rate (Badenes and Maoz, 2012). It could be possible that sub-Chandrasekhar mergers could lead to SNe Ia, or perhaps single-degenerate scenarios are more common than previously thought. An alternative model, proposed by Kushnir et al., 2013, suggests that direct collisions of WDs in triple systems,

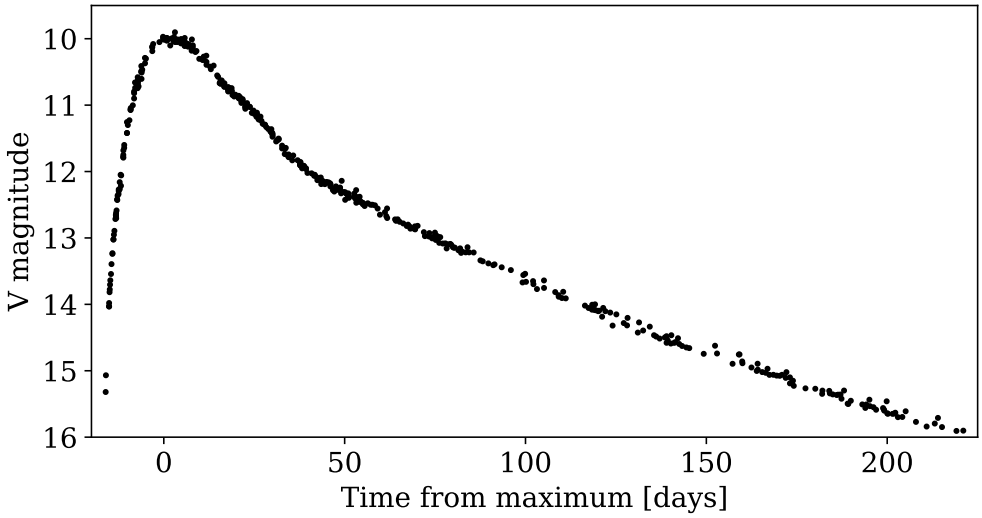

FIGURE 1.4: Photometric lightcurve of SN 2011fe, a type Ia supernova, in the V-band. Note the change in slope around 50 days post-maximum, corresponding to a change in the dominant decaying element. Data from Guillochon et al. (2017).

by which the third companion perturbs the two WDs to infall radially, actually constitutes the majority of type Ia events and can explain the observed SNe Ia rate.

Other WD-WD merger events may be more subtle than the blazing anomaly SN 2003fg, and it is possible the two WDs may even be of different composition, e.g. a C-O and/or a O-Ne-Mg WD. As such, we require improved methods of identifying these outliers so that we may screen them out of cosmology studies.

1.2 Optical Spectrum

All supernovae are classified based on their spectra. A representative example of a spectrum for each type of supernova is shown in Figure 1.5. Type I supernovae are defined by the absence of hydrogen emission lines, whereas type II supernovae are defined by exhibiting strong hydrogen features. The subclass type Ia is defined by the presence of a (blue-shifted) Si II λ6355 absorption feature at 6150 Å near maximum light. The presence of helium lines define the other two subclasses type Ib and Ic. Figure 1.6 provides a decision tree for classifying a supernova's type based on its spectrum.

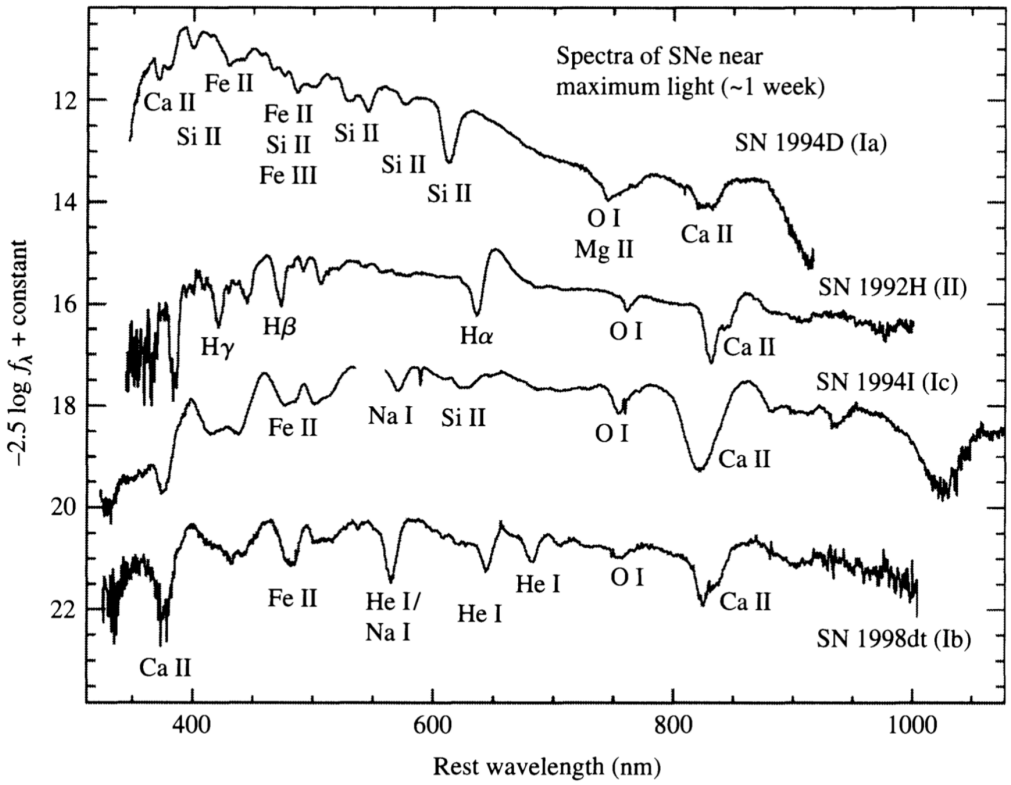

FIGURE 1.5: Example spectra for the different supernova types. From Carroll and Ostlie (2007).

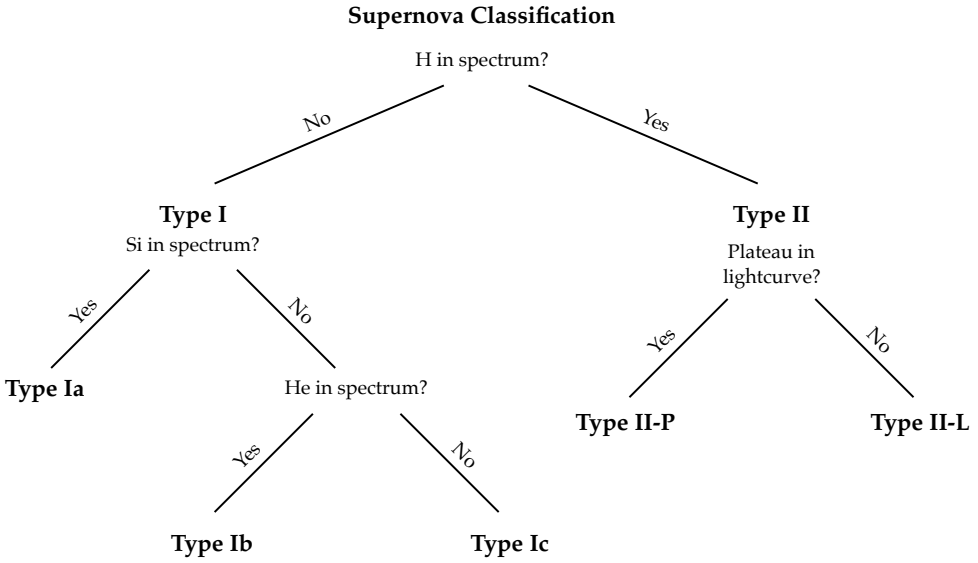

FIGURE 1.6: A decision tree showing the classification of SNe types based on the presence of various elements in their spectra. Reproduced from Carroll and Ostlie (2007).

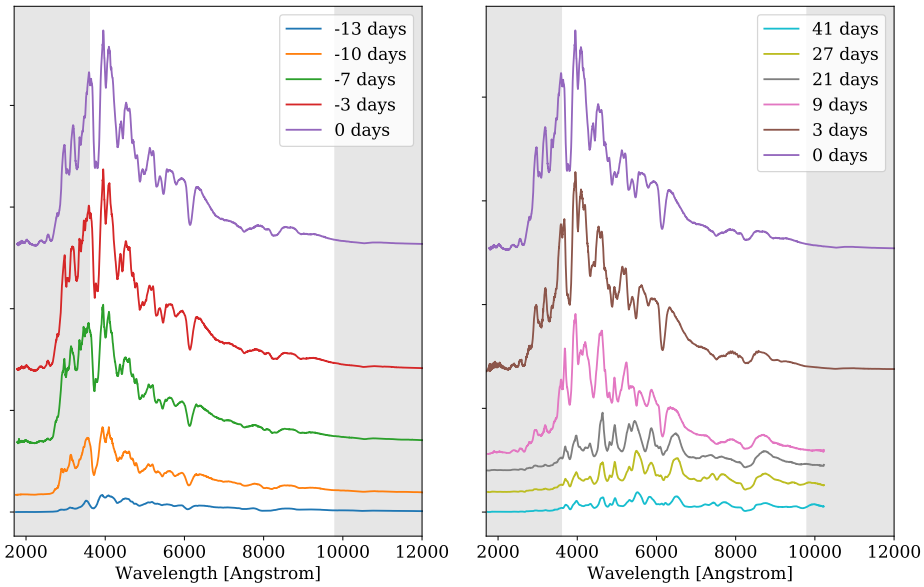

FIGURE 1.7: Spectroscopic observations of SN 2011fe, a type Ia supernova, from pre-maximum brightness (left) to post-maximum (right), shifted vertically for visibility. The non-shaded region indicates the wavelength regime of DESI. Data from Guillochon et al. (2017).

The evolution of a SN Ia's luminosity, i.e., its *light curve*, is driven by radiation produced in the ^{56}Ni $\rightarrow$ ^{56}Co $\rightarrow$ ^{56}Fe radioactive decay chain (Truran, Arnett, and Cameron, 1967; Colgate and McKee, 1969). Figure 1.4 shows an example of a SN Ia lightcurve. As the radiant photons come from this radioactive decay process, the luminosity follows an exponential decay according to the decay rate of the elements. Since the decay rate is a unique value for a given element, measuring the slope of the supernova lightcurve yields the elemental abundance. This is particularly useful since ^{56}Ni dominates the peak SN Ia luminosity, and combined with the spectrum the fractional mass of ^{56}Ni of the progenitor can be determined. Thus, determination of the ^{56}Ni mass can help constrain the detonation mechanism, as in the case of SN 2003fg in which a super-Chandrasekhar mass progenitor was discovered.

The spectrum of a SN Ia is characterized by strong absorption and emission lines superimposed on a thermal blackbody continuum. Early-time spectra (from the moment of explosion to around maximum brightness) contain many Mg II, Si II, S II, Ca II, and O I features, as well as some weak Fe II and Fe III lines. The strongest of these are the the broad Si II λ6355 absorption feature as well as the Ca II H&K $\lambda\lambda$3934, 3968 lines.

After about two weeks from maximum brightness, the ejected material expands beyond the photosphere and ^{56}Co $\rightarrow$ ^{56}Fe decay becomes the dominant source of luminosity. The spectrum becomes dominated by permitted emission lines of Fe II, as well as some intermediate-mass elements like Si II and Ca II. At this stage the changes in spectral features become more gradual. Late-time spectra ($\geq$ 4 months, also called the *nebular phase*) show a variety of blended forbidden Fe II, Fe III, and Co III transitions. Figure 1.7 shows the evolution of the optical spectrum of the archetype SN Ia SN 2011fe. For a more comprehensive overview of type Ia supernova spectra, see Filippenko (1997) and Parrent, Friesen, and Parthasarathy (2014).

Blue-shifted spectral features yield mean ejecta velocities of ~5,000 km/s and peak velocities exceeding 20,000 km/s, about 1/15th the speed of light. These velocities are consistent with the energy released from fusing one Chandrasekhar mass of carbon and oxygen into the intermediate-mass (Si or Ca) and heavy (Ni, Fe) elements observed in the spectra (Hillebrandt and Niemeyer, 2000). Spectra thus not only provide information about what elements are present, but also provide constraints on the explosion kinematics.

Both the optical spectra of SNe Ia and their photometric lightcurves are, in general, quite homogenous across observations when accounting for the time-from-maximum (i.e., the *epoch*). As a result, a great deal of effort has gone into creating model templates of lightcurves and spectra alike, for example those developed by Hsiao et al. (2007). Agreement, or even disagreement, between these templates, observations, and theorized explosion mechanisms is an important test in understanding the astrophysics of SNe Ia. Deviations from this homogeneity can also provide clues to the exact nature of the SN Ia progenitor and host environment.

1.3 Motivation for Spectroscopic Searches

SNe Ia have a wide range of astrophysical utilities (see Howell, 2011). Most famously, they play a crucial role in the so-called *distance ladder* of obtaining distances to galactic and extragalactic objects. Because of the uniformity of the physics governing an object consisting of roughly one Chandrasekhar mass of degenerate carbon and oxygen, the

resulting supernova produces a common explosive yield. This translates into a common intrinsic luminosity for SNe Ia, and permits their use as *standard candles*. The absolute magnitude of a SNe Ia has been found to be (Hillebrandt and Niemeyer, 2000)

$$M_B \approx M_V = -19.30 \pm 0.03. \tag{1.1}$$

With this absolute magnitude representative of a "normal" SN Ia, a measurement of the apparent magnitude immediately yields the distance to the SN Ia. Using this method to determine the distance to 10 SNe Ia in host galaxies with redshifts between 0.16 and 0.62, Riess et al. (1998) were able to show that the expansion of the universe was accelerating, a discovery worthy of the 2011 Nobel Prize in Physics. A more recent study using this method has tightened the uncertainties on the Hubble constant to just 2.4% (Riess et al., 2016). To further improve upon the accuracy of these cosmology studies, a large sample of SNe Ia is needed. In addition, spectroscopic measurements are crucial for calibrating photometric observations and for providing complementary redshift coverage.

Spectra of SNe Ia contribute a plethora of valuable science. Near maximum light, spectra are how SNe are classified, so understanding the range of features in a SN spectrum is critical to understanding the physical mechanism behind it. Spectra provide constraints on the energetics of the explosion by inferring explosion velocities from the blueshift and redshift of absorption lines relative to the host galaxy. The diversity of elements created via nucleosynthesis during the carbon burning and explosion phases of the SNe reveal themselves through their characteristic emission and absorption features, which are dependent on the composition and mass of the progenitor object. Identifying the progenitor can provide necessary screening of SNe Ia that may systematically deviate from the "standard candle" relation, such as a super-Chandrasekhar mass progenitor.

All of these science cases require a large and diverse population of SNe Ia to be exhaustively studied, both photometrically and spectroscopically. Hence, the pursuit of novel means of detecting and classifying these objects is strongly motivated, and needs to be done automatically to analyze SNe Ia spectra using the next generation of spectroscopic surveys. In this thesis I present and discuss several approaches towards this task.

Chapter 2

DESI

2.1 Overview of the Survey

The Dark Energy Spectroscopic Instrument (DESI) is a next-generation multi-fiber spectrograph that will compile the largest spectroscopic database of galaxies to date. DESI will be operating from the 4-meter Mayall telescope (see Figure 2.1) at Kitt Peak National Observatory in Tucson, Arizona. First light is scheduled for September 2019. DESI will build on the success of the Sloan Digital Sky Survey (SDSS) and Baryon Oscillation Spectroscopic Survey (BOSS) dark energy experiments and precursor large-sky spectroscopic surveys. Surveying 14,000 deg^2 of the night sky over five years, DESI will measure spectra and redshifts for 25 - 35 million galaxies to create the largest and most accurate

FIGURE 2.1: The Mayall 4-m telescope dome (left) and telescope (right), where the DESI experiment will be conducted. Credit: NOAO.

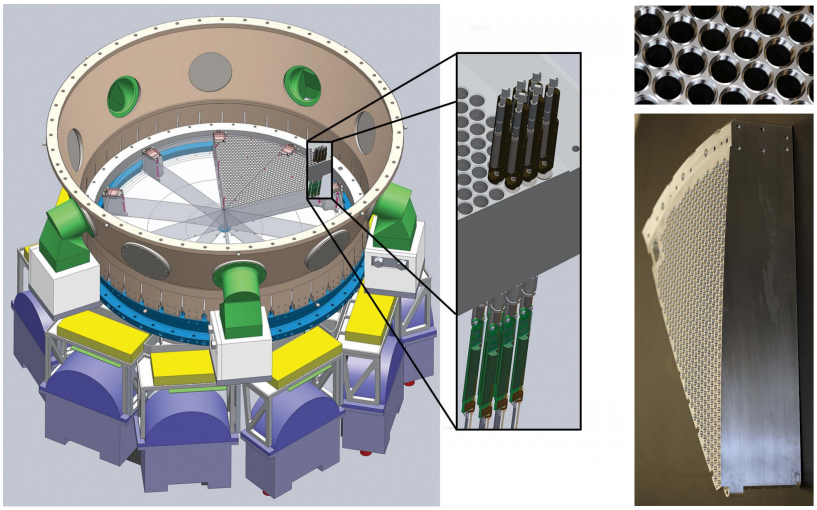

FIGURE 2.2: The DESI imaging array: Ten "petals" (right) contain 500 robotic fiber-optic positioners that each feed a three-armed spectrograph (DESI Collaboration et al., 2016b).

3D map of the Universe and provide new constraints on the dark energy equation of state (DESI Collaboration et al., 2016a).

The DESI instrument consists of 5,000 robotically-actuated optical fibers that feed ten three-arm spectrographs (see Figure 2.2). At each arm of the spectrograph is a 4k×4k CCD camera: The blue arm (b) operates at 3600 - 5930 Å, the red arm (r) at 5660 - 7720 Å, and the near-IR arm (z) at 7470 - 9800 Å. The full wavelength coverage for DESI is thus $\lambda \in [3600\ \text{Å}, 9800\ \text{Å}]$. Each of the 5,000 optical fibers can be robotically positioned within the focal plane to point at the location of a target galaxy. In principle, a single exposure can therefore collect spectra for 5,000 objects simultaneously within the instrument's 8 deg^2 field-of-view (DESI Collaboration et al., 2016b).

2.2 Time-Domain Science with DESI

While DESI's primary science goal is to measure cosmic expansion via observations of the Baryon Acoustic Oscillation (BAO) peak, DESI provides an excellent experimental setting to observe interesting objects spectroscopically. DESI will collect spectra for luminous red galaxies (LRGs) up to $z = 1.0$, emission line galaxies (ELGs) up to $z = 1.7$, and quasi-stellar objects (quasars, or QSOs) for $2.1 < z < 3.5$. The Bright Galaxy Survey

(BGS) will cover 10 million galaxies with $z \leq 0.4$. With this large sample of objects, DESI can search for interesting time-dependent phenomena such as supernovae at $z < 0.1$, and gravitationally lensed supernovae at higher redshifts (DESI Collaboration et al., 2016a).

The overlap between DESI and other large imaging surveys, such as the Dark Energy Survey (DES), Large Synoptic Sky Telescope (LSST), and Zwicky Transient Facility (ZTF), allows for interesting complementary science. DESI can identify known classes of transients and conduct follow-up spectroscopy of candidates discovered by these imaging surveys, and also holds the potential to discover new classes of objects. For instance, DESI is expected to observe 10^5 SNe in the BGS, and potentially just as many lensed supernovae in the LRG sample at higher redshifts. Obtaining redshifts for tens of thousands of SNe host galaxies will allow for high precision distance measurements in collaboration with the SNe samples from LSST and DES (DESI Collaboration et al., 2016a).

Supernova studies have historically been limited by the low statistics of spectroscopic observations. Logistical difficulties from impromptu follow-up of transients, incomplete wavelength and epoch coverage (e.g. multiple experiments operating at different wavelength ranges contributing spectra of the same object at different epochs), low signal-to-noise ratios, and spectroscopic line-blending have all posed challenges towards obtaining quality spectroscopic measurements of supernovae (Parrent, Friesen, and Parthasarathy, 2014). Spectra obtained by DESI, as well as by follow-up from other surveys, will be essential in developing comprehensive studies of SNe Ia and other interesting transients.

Chapter 3

Dataset

3.1 DESI Simulation Pipeline

To develop and test various methods of identifying SNe Ia in galaxy spectra, we produce a mock dataset using the DESI Collaboration's simulation pipeline, `desisim`[1]. The simulation builds Monte Carlo spectra of the various object types (LRGs, QSOs, ELGs, BGS, FSTD (metal-poor, main sequence turnoff standard stars), sky, etc.) based on template spectra derived from spectral models used for fitting SDSS/BOSS galaxy redshifts. By default, the pipeline simulates the BGS in the redshift range $0.01 \leq z \leq 0.4$, LRGs in the range $0.5 \leq z \leq 1.0$, ELGs for $0.6 \leq z \leq 1.6$, and QSOs for $0.5 \leq z \leq 4.0$, although the user can define their own desired redshift ranges. The redshift distribution for our simulated sample is shown in Figure 3.1.

The pipeline first simulates the details of the exposure, including a range of observational conditions including moon fraction, airmass and exposure time. The Monte Carlo spectrum is then folded through a detector response function for DESI, which attempts to realistically model the effects of the various components of the instrument and environment on the measured spectrum. For example, fiber losses are modeled by taking into account the source's 2D spatial profile on the sky, the transformation of sky coordinates to physical coordinates on the fiber at the focal plane (i.e. a round fiber being projected as an elliptical aperture on the sky), and the atmospheric and instrumental point

[1] Available at https://github.com/desihub/desisim

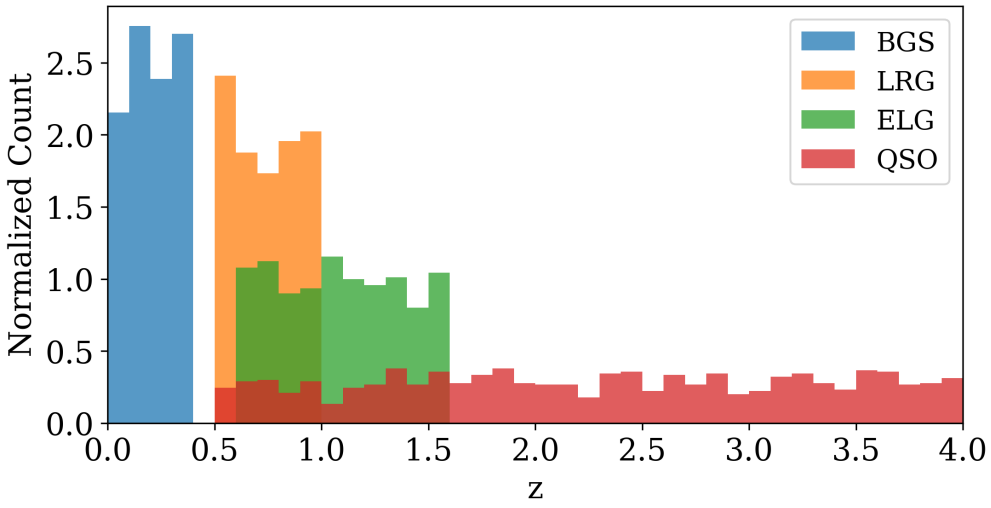

FIGURE 3.1: Redshift distribution for the our mock catalog of galaxies.

spread function (PSF). Fiber throughput and various sources of noise are also handled accordingly.

The simulation we use, encapsulated in the utility `QuickGen`, considers a one-dimensional 'smearing' of the input spectra across the detector. A full, and more realistic, two-dimensional smearing is available, but is 2-3 orders of magnitude slower to run. The 1-D version still models the relevant instrumental effects and is sufficiently realistic for our needs, and therefore is used for producing our mock dataset. When generating a collection of observations (e.g. an exposure including all 5,000 fibers), `desisim` also sets aside 10% of the fibers for calibration: 8% simulate standard stars and 2% simulate sky exposures, which are used to remove atmospheric emission and absorption lines from the extragalactic spectra.

The final, flux-calibrated spectra are saved in FITS format for each camera and spectrograph. For example, a simulation consisting of 5,000 objects of 'BGS' type would result in a total of 4,500 BGS galaxies, 400 calibration stars, and 100 sky exposures. The 5,000 fibers are divided into into the 10 spectrographs, where each spectrograph 'reads' 500 of the fibers. An exposure of a spectrum for each object is then generated for the b, r, and z cameras. The final result is 30 .FITS files: 1 per spectrograph and 3 per camera, each containing 500 simulated spectra.

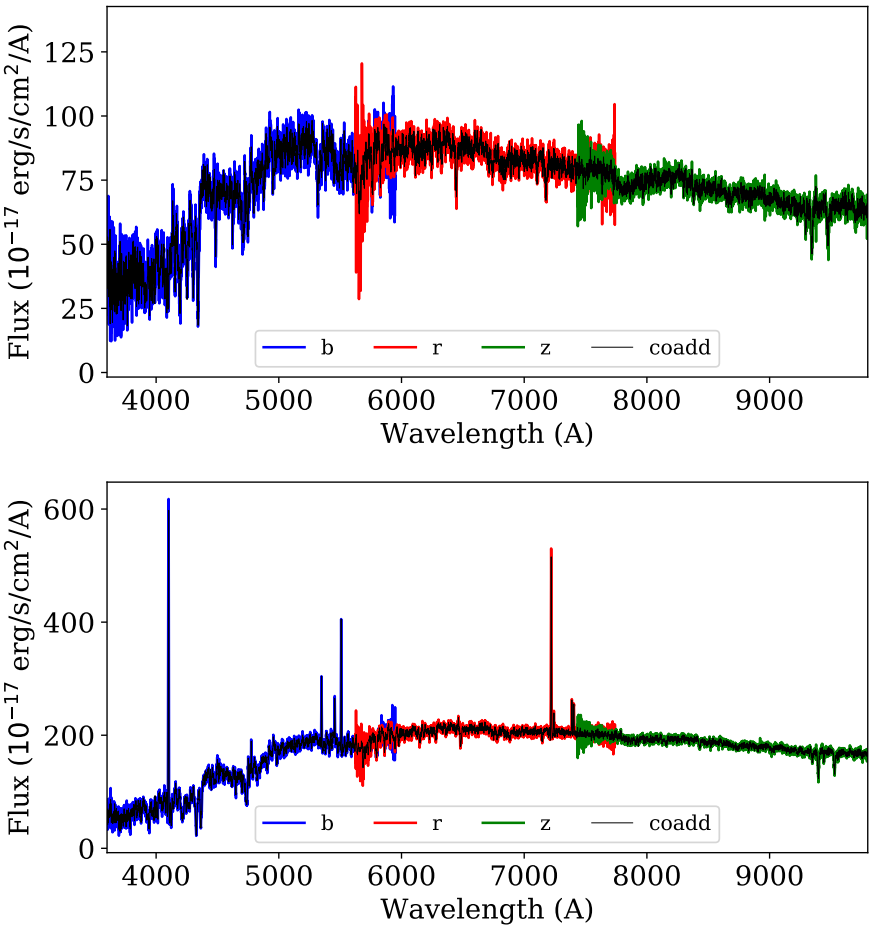

FIGURE 3.2: Examples of the coaddition of the b, r, and z camera exposures for two simulated BGS galaxies.

To obtain the final spectrum for each object, the exposures from the b, r, and z cameras are coadded into one spectrum covering the full wavelength range. The coaddition algorithm combines the flux in the overlapping wavelength regions of the three cameras, accounting for the uncertainties in the flux values. Two examples of this process are shown in Figure 3.2 for simulated BGS galaxies.

The coaddition step is the most expensive part of the simulation process. A single b, r, and z coadd can take upwards of 1 - 5 minutes. However, the process of coadding these exposures after they have been generated by `QuickGen` (which takes around 5-10 minutes to create 5,000) is highly parallelizable. Therefore, we utilize the BlueHive[2] computing cluster at the University of Rochester to produce a large mock dataset with 10,000 coadded

[2] http://www.circ.rochester.edu/

spectra produced approximately every hour. The final mock catalog consists of 18,900 BGS galaxies, 900 LRGs, 900 ELGs, 900 QSOs, and a corresponding number of star and sky exposures. This analysis focuses on the BGS dataset, since that is where we will be most likely to observe a large number of supernovae.

3.2 Type Ia Supernova Injection

To create composite spectra of galaxies containing supernovae, we inject SN Ia template spectra into each simulated galaxy to produce an additional set of spectra. This set is labeled "has SN" and will be used to test our SN Ia detection algorithm in Chapter 4.

The SN Ia templates we use are the collection of observations of the well-observed type Ia supernova SN 2011fe from the Open Supernova Catalog (Guillochon et al., 2017). We select only the spectra that sufficiently cover the target wavelength range (must have a maximum wavelength of at least 9800 Å, and minimum wavelength of at most 2500 Å). This guarantees that, even at the highest redshift in the BGS sample ($z = 0.4$), the template will cover the full DESI wavelength range. Further, any observations for which the flux unit is defined as "uncalibrated" are discarded. This leaves us with 10 template spectra ranging from 13 days pre-maximum brightness to 41 days post-maximum brightness. These template spectra are shown in Figure 1.7.

The process of injecting a SN Ia into a galaxy spectrum is as follows. A random epoch from the light curve is chosen using a uniform prior, and the corresponding spectrum is selected for injection. The SN Ia template is redshifted to the redshift of the host galaxy, and is interpolated onto a new wavelength grid corresponding identically to that of the host galaxy. The resampled, redshifted SN Ia template is then scaled to the host galaxy and is superimposed by adding the SN Ia and galaxy flux values at each wavelength bin. The final composite spectrum is then written to a new FITS file with identical structure to that of the galaxy, but with the "has SN" flag changed from "False" to "Ia".

We consider two different approaches when scaling the SN spectrum to the host galaxy. The first is a relative scaling by median flux ratios. This scaling reflects the relative continuum brightness between a typical SN Ia and its host, without being biased by the

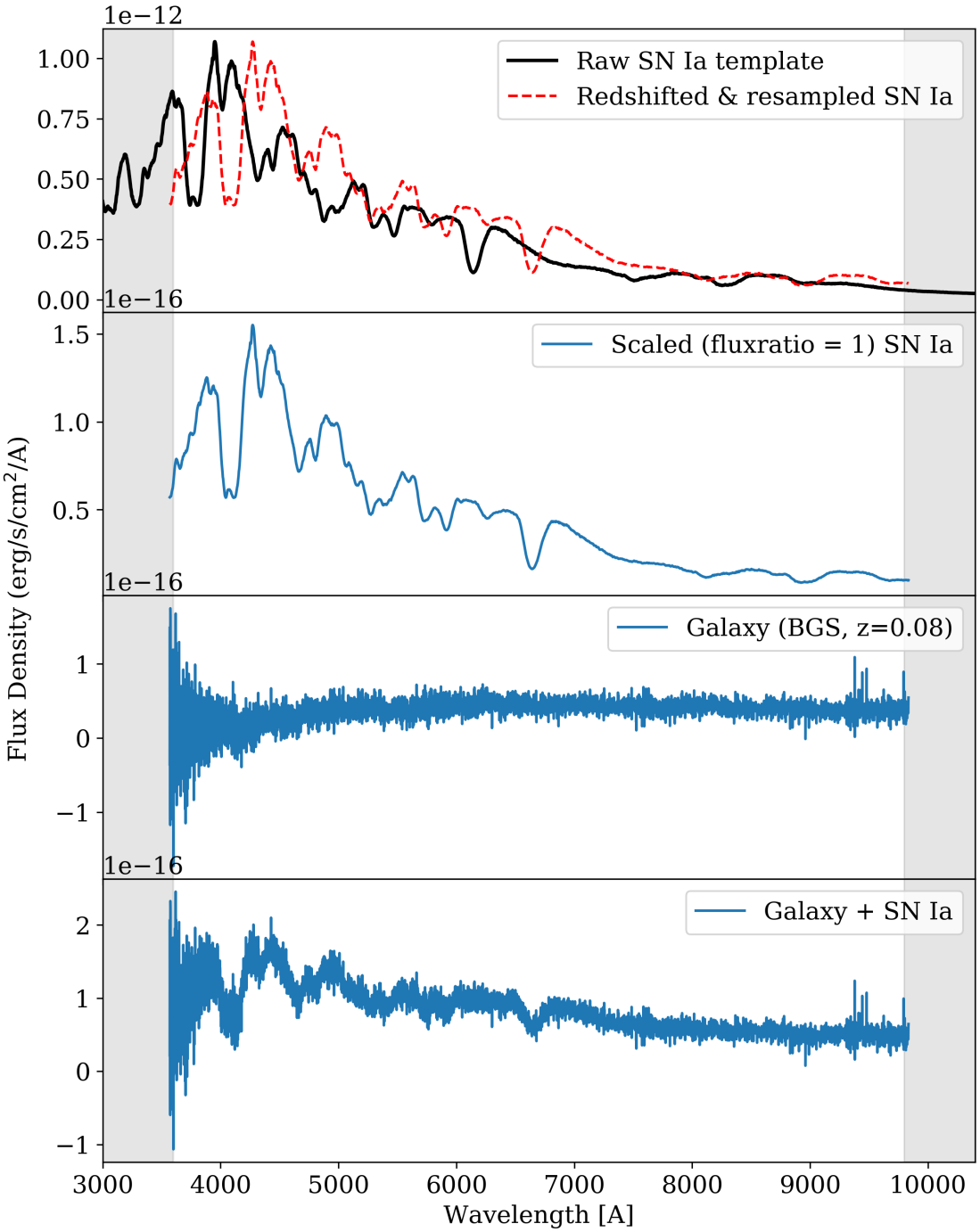

FIGURE 3.3: An example of the SN Ia template spectrum at peak brightness being injected into a simulated BGS galaxy spectrum, with a scaling of fluxratio = 1. The non-shaded region indicates the wavelength regime of DESI.

strong emission lines in the host spectrum. The scaling is given by

$$f_{SN} = f \cdot \frac{\text{median}(f_{galaxy})}{\text{median}(f^0_{SN})} \cdot f^0_{SN}, \tag{3.1}$$

where f^0_{SN} is the flux of the SN Ia template, f is called the "fluxratio," and f_{SN} is the scaled flux. Using this method, we generate four complete samples each containing 18,900 SN Ia + galaxy composite using the simulated BGS dataset, for fluxratio = 1 (the "bright" sample), 0.5 (the "dim" sample), 0.1 (the "faint" sample), and 0.01 (the "very faint" sample).

The second approach is uses an accurate distance scaling to preserve the "standard candle" nature of a SN Ia. First, the luminosity distance to the galaxy is determined by its redshift z, where

$$D_L = (1+z)D_{CM}, \tag{3.2}$$

where D_{CM} is the comoving distance given by (assuming a flat cosmology)

$$D_{CM} = (c/H_0)\int_0^z \frac{1}{E(z)}dz, \quad E(z) = \sqrt{\Omega_M(1+z)^3 + \Omega_\Lambda}, \tag{3.3}$$

where $c = 299,792,458$ m/s is the speed of light, H_0 is the current value of the Hubble Constant, Ω_M is the matter energy density, and Ω_Λ is the dark energy density of the universe. We use $H_0 = 73.24$ (Riess et al., 2016), $\Omega_M = 0.31$ (Planck Collaboration et al., 2016), and therefore $\Omega_\Lambda = 0.69$. Then, we may scale the SN Ia flux by

$$f_{SN} = \left(\frac{D_{L,\,SN}}{D_{L,\,galaxy}}\right)^2 \cdot f^0_{SN}, \tag{3.4}$$

where $D_{L,\,SN}$ is the luminosity distance to SN 2011fe, which is provided in the source data from the Open Supernova Catalog. We use this method to create an additional sample of 18,900 SN Ia + galaxy composite spectra, using the simulated BGS dataset. Thus, for each of the 18,900 BGS galaxies, there is a corresponding "bright," "dim," "faint," "very Faint," and distance-scaled SN Ia host. Examples of the five SN Ia host datasets are provided in Appendix A.

Chapter 4

Techniques for Identifying Type Ia Supernovae

In this chapter I discuss various methods of identifying galaxy spectra that contain a SN Ia. Several approaches to SN Ia identification that have been used in past studies are highlighted, and novel methods developed in this thesis are presented. We find that a simple set of defined features can be used to identify SNe Ia without modeling or fitting the emission from the host galaxies.

4.1 Principal Component Analysis

The traditional approach to detecting any form of astrophysical transient is by repeated imaging of the sky. By studying a particular region over many nights, one can subtract the "usual" emission accumulated over "ordinary" nights to reveal any excess emission that might occur over a short period of time. Follow-up imaging and spectroscopy can then help determine the identity of the candidate transient.

In general, spectroscopic observations involve a greater observational cost than photometric imaging. Because of this, spectroscopic surveys are less efficient and cannot cover the same regions of the sky as quickly as their photometric counterparts. This makes subtraction of previous epochs of data more difficult, and is thus not an efficient means of searching for transients.

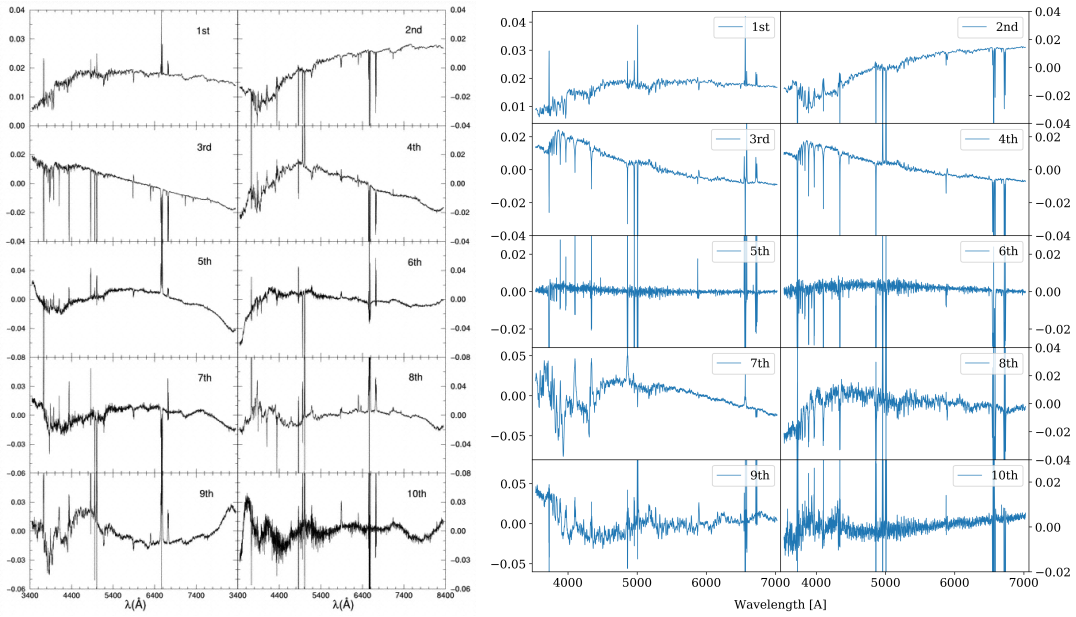

FIGURE 4.1: (Left) First 10 KL eigenspectra of 170,000 SDSS galaxy spectra (Yip et al., 2004). (Right) First 10 eigenspectra using 18,900 simulated BGS galaxies from the `desisim` pipeline.

Instead, one can develop a spectroscopic model of a typical galaxy, and subtract this model from galaxies observed spectroscopically. Emission not due to physical properties arising in such a "typical" galaxy will thus be revealed. One such method of modeling galaxy spectra is by principal component analysis (PCA) of a large sample of galaxies. PCA involves decomposing the vector space composed of spectra into principal components that can then be used to span the original space. The components, or "eigenspectra," can then be used to fit any given galaxy by a linear combination.

Yip et al. (2004) develop such a classification using galaxy spectra from SDSS, using a Karhunen-Lov (KL) transformation to obtain the eigenbasis. Their approach, which we follow in our analysis, is as follows. Each galaxy spectrum in the SDSS catalog is thought of as an axis in a multidimensional space, where the entire set of galaxies comprises the full space and represents an $M \times N$ matrix, where M is the number of galaxies and N is the number of wavelength bins in the spectrum. Let $f_{\lambda_k i}$ be the flux in wavelength bin λ_k for the i^{th} galaxy (i.e. the i^{th} galaxy's spectrum), where $i \in [1, M]$ and $k \in [1, N]$.

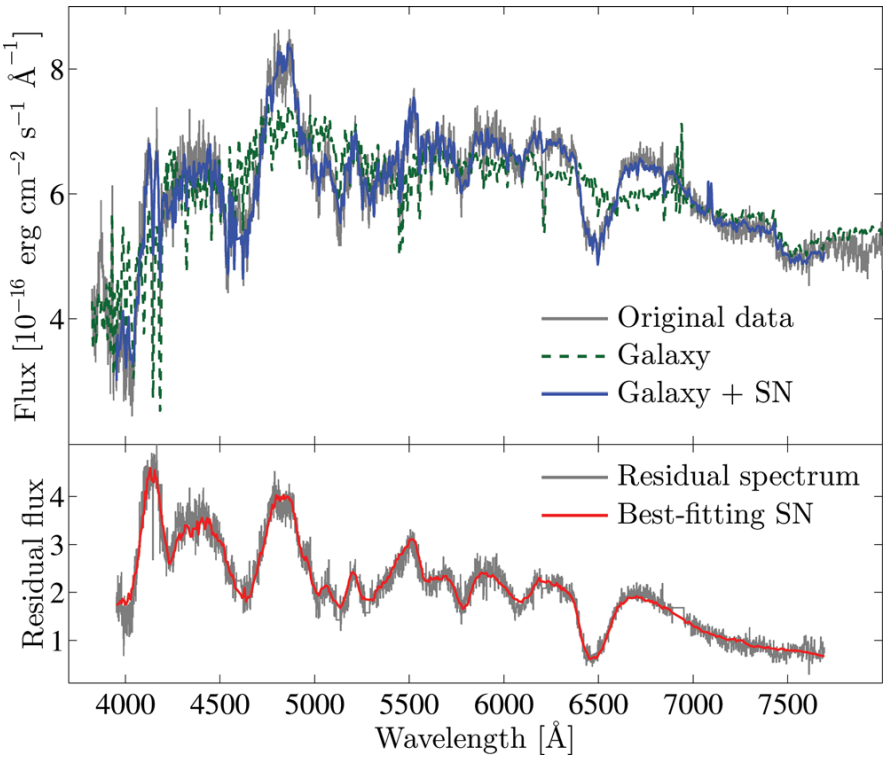

FIGURE 4.2: Example SN Ia discovered in the SDSS database (Graur and Maoz, 2013). The green dashed line gives the fit using the 10 KL eigenspectra from Yip et al. (2004).

Then, define a correlation matrix

$$C_{\lambda_k \lambda_l} = \sum_{i=1}^{M} \hat{f}_{\lambda_k i} \hat{f}_{i \lambda_k} \tag{4.1}$$

between each pair of normalized spectra $\hat{f}_{\lambda_k i}$. That is,

$$\hat{f}_{\lambda_k i} = \frac{f_{\lambda_k i}}{\sqrt{f_{\lambda_k i} \cdot f_{\lambda_k i}}}. \tag{4.2}$$

Now, the eigenspectra may be obtained by solving the eigenvalue equation

$$\mathbf{U}^{\mathrm{T}} \mathbf{C} \mathbf{U} = \mathbf{\Lambda}, \tag{4.3}$$

where $\mathbf{U}$ is a matrix whose i^{th} column contains the i^{th} eigenspectrum, denoted $e_{i\lambda_k}$, and the matrix Λ is a diagonal matrix containing the eigenvalues of the correlation matrix.

n	CEV	n	CEV	n	CEV	n	CEV
1	96.55%	6	99.46%	11	99.51%	16	99.52%
2	98.69%	7	99.47%	12	99.51%	17	99.53%
3	99.09%	8	99.49%	13	99.52%	18	99.53%
4	99.33%	9	99.50%	14	99.52%	19	99.53%
5	99.40%	10	99.50%	15	99.52%	20	99.53%

TABLE 4.1: Cumulative explained variance (CEV) of the eigenspectra for the simulated BGS dataset, for n components.

This may be done numerically with a singular-value decomposition (SVD) algorithm, for example with the `numpy.linalg.svd` function in Python. The output produces a list of eigenspectra that may then be used to 'fit' any given galaxy by a simple linear combination:

$$f_{\lambda_k i} = \sum_{j=1}^{M} a_j e_{j\lambda_k}, \tag{4.4}$$

where m is the total number of eigenspectra. Yip et al. (2004) find that the first three KL eigenspectra explain 98% of the total variance in the galaxy dataset, and the first 10 contain 98.52% of the total variance. Figure 4.1 show what the first 10 eigenspectra look like.

The approach of using the eigenspectra to fit and subtract the galaxy from a given spectrum is used by Krughoff et al. (2011) and Graur and Maoz (2013) to search for SNe Ia in the residual spectra. Both studies use a series of supernova templates to fit to the residuals in search of the best-fit supernova. An example of a galaxy hosting a SN Ia discovered by Graur and Maoz (2013) is shown in Figure 4.2. The spectrum is fit much more strongly by a model containing the galaxy eigenbasis + SN template than a model containing only the eigenspectra. Graur and Maoz (2013) find a total of 90 SNe Ia among 700,000 SDSS galaxies using this technique.

We use a similar approach with eigenspectra built from the simulated BGS dataset, containing 18,900 galaxy spectra. To define an appropriate $M \times N$ matrix to represent the space of galaxy spectra, all spectra are interpolated onto a common wavelength grid ranging from the minimum wavelength of the lowest redshifted galaxy to the maximum wavelength of the highest redshifted galaxy. This ensures the maximum wavelength coverage such that all galaxies are on the same wavelength grid with defined fluxes. For our

entire analysis, this common wavelength grid is used. The first 10 eigenspectra built from BGS galaxies with no injected SN Ia are shown in Figure 4.1. Note the similarity between the first three eigenspectra - an indication of the quality of the realism of the simulation. In our sample, the first 3 eigenspectra explain 99% of the variance in the simulated BGS dataset, and the first 10 explain 99.5% of the variance. A table showing the cumulative explained variance for a given number of eigenbasis components is given in Table 4.1.

In the next section, we apply a similar residual search to that used by Graur and Maoz (2013) and Krughoff et al. (2011) (using the eigenspectra to model the galaxy), but instead of fitting supernova templates we use a machine learning predictor.

4.2 Random Forest Classifier

Given a dataset x_i containing N observations, each with class label c_i and feature set f_{ij} containing M features, we wish to classify the data x_i according to their class labels c_i, given the values of the particular features f_{ij} for each x_i. In our case, the x_i are the galaxy spectra, the c_i are a flag indicating if the galaxy hosts a SN Ia or not, and the features f_{ij} are to-be-defined values derived from the data x_i, chosen in a way to aid in the classification. To perform this classification, we use a random forest.

A random forest (RF) is an ensemble learning method often used for regression and classification tasks that utilizes decision tree classification. A decision tree is an object consisting of a number of nodes, where each node contains "branches" that split the path of a given input data object on some criterion (e.g. the value of a particular "feature" in an observation). Figure 1.6 is an example of a decision tree. The path through the decision tree that an observation takes is terminated by a "leaf" node that represents the class label of the observation. A RF is constructed from an ensemble of decision trees (hence the term *forest*) that are each built using randomized criteria for splitting (hence the term *random*).

The quality of a given split is measured using the Gini impurity, which measures the tree's probability of having incorrectly labelled a object at that node. This is determined

by calculating the fraction f of items reaching that node with a given label, and multiplying by the probability of that node misclassifying it, (1 - f). f can be found by simply summing all instances of that label in that node and dividing by the total number of items reaching the node. The quantity $f(1-f)$ is summed over all the labels that have reached that node to give the Gini impurity.

The RF is constructed from a "training" dataset with known *a priori* class labels. A single tree is grown by drawing with replacement from a random subset of the training sample, and using a random subset of the feature set to develop the splitting criteria. Each tree in the forest is grown (i.e. splitting nodes/adding branches) until either all leaves are pure (all samples that reach that leaf have the same label, giving a Gini impurity of zero) or until all leaves contain just one sample. In either case, every leaf node represents a unique class label. This is repeated N_{tree} times, where N_{tree} is the number of trees in the forest.

The trained RF predicts the class label of a new item by propagating it through all the trees in the forest and averaging the resulting class labels to give the overall predicted class. Due to this averaging, RF predictions in general have low variance and are robust with respect to noise (Breiman, 2001). Further, since each tree in the forest can "vote" for which class it thinks the input observation belongs to, the final prediction can be given in terms of a probability. For example, if the 70% of the N_{tree} trees vote for class A, 20% for class B, and 10% for C, the observation could be assigned to "class A with probability 0.7, class B with probability 0.2, and class C with probability 0.1." The overall predicted class is that which wins the majority, in our example this would be class A.

Another useful feature of a RF is that it can rank the features used for classification by their relative importance. This is done naturally during the training stage by summing the gini impurity for a node's children, and subtracting this from the gini impurity of the node itself, across all trees in the forest. Splitting on a relatively "important" variable will thus result in a large decrease in the impurity, whereas a relatively "unimportant" variable will not cause much of a change after splitting. These are automatically weighted across all features and can thus be used as a relative importance measure for the feature

set. This is a powerful way to remove unimportant features and learn which are most useful for classification purposes.

An important test of a predictive machine learning algorithm, especially one involving supervised learning (training with known class labels), is to ensure one has not *overtrained* their classifier. Overtraining occurs when the classifier "overfits" the training data. That is, the criteria on which the classifier determines the class of a given data object is highly specialized to the particular subset of data used for training. In this case, the RF has a perfect or near-perfect ability to predict the class of objects in the training set (that it has already seen), but may fail at predicting the class of unseen data. Overfitting generally results from using more parameters than degrees of freedom in the data, and thus any statistical fluctuations in the particular choice of training data can be fit, disrupting the predictive power of the classifier.

To avoid overfitting, we use a Monte Carlo cross-validation of our RF. This works by partitioning the full dataset into a training set, containing 75% of the data, and a testing set, containing 25% of the data. Which values are partitioned into each set is chosen at random. The RF is then trained using the training sample and applied to the test sample to gauge its predictive ability. This process is repeated N_{cv} times, where N_{cv} is chosen to be sufficiently large that each data point has a high probability of being in both the training and testing set at some iteration. The results of the predictions on the test sample are averaged over the N_{cv} cross-validations to provide an estimate of the variance in the RF's predictive ability.

Our approach to spectroscopically identifying SN Ia involves training a RF to classify spectra as either having or not having a SN Ia present, using a set of features derived from the spectrum. For our RF classifier, we use the functions provided in `sklearn.ensemble.RandomForestClassifier` from the Python library *scikit-learn*. By default, the RF is built from 10 decision trees. In the following subsections, we examine two different means of deriving features from the spectrum.

Feature id	Dominating Line	Feature Range [Å]	Lower Region [Å]	Upper region [Å]
f1	Ca II H&K	3700 - 3900	3450 - 3800	3800 - 4100
f2	Si II λ4000	3950 - 4100	3800 - 3950	4000 - 4200
f3	Mg II λ4300	4200 - 4400	3850 - 4250	4300 - 4700
f4	Fe II λ4800	4700 - 5000	4300 - 4700	4950 - 5600
f5	S II W	5300 - 5600	5050 - 5300	5500 - 5750
f6	Si II λ5800	5700 - 5900	5400 - 5700	5800 - 6000
f7	Si II λ6150	6000 - 6300	5800 - 6100	6200 - 6600

TABLE 4.2: SN Ia spectral features from Nordin et al. (2011).

4.2.1 Features Using Residual Spectrum

Our first approach involves searching for features in the residual spectrum, after subtracting the fit using the first 10 eigenspectra generated from the BGS sample, using the same common wavelength grid that was used to define the eigenspectra. An example of this subtraction for a bright SN Ia hosting galaxy and a normal galaxy is shown in Figure 4.3. As can be clearly seen, the eigenspectra fit the emission due to the galaxy very well, and the residual for the normal galaxy is nearly feature-less. The eigenspectra fail to fit features due to the SN Ia in the hosting galaxy, and these broad absorption features are strongly visible in the residual spectrum.

To define a feature set, we take the broad absorption features used by Nordin et al. (2011) to characterize spectral properties of SN Ia out to $z \sim 0.3$. Since the BGS sample extends to $z \leq 0.4$, these features will be visible in all rest-frame shifted BGS galaxies.

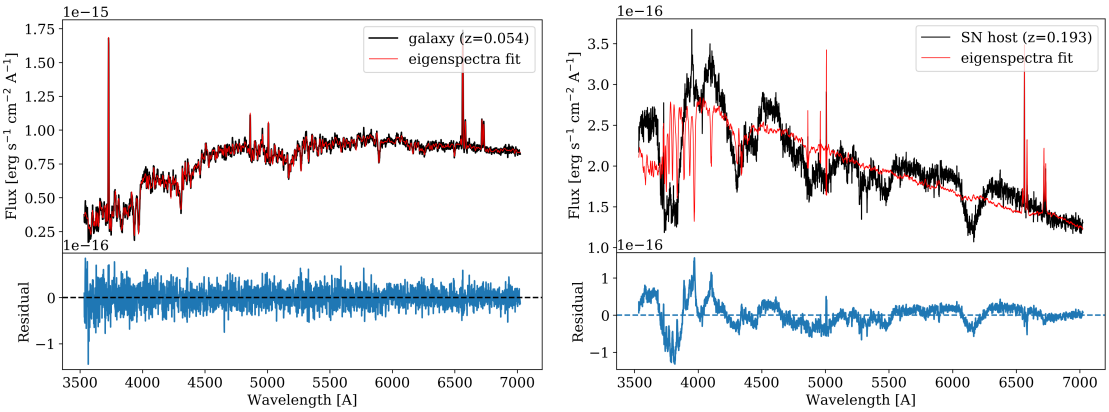

FIGURE 4.3: Eigenspectra fit and subtraction for a regular BGS galaxy (Left) and SN Ia hosting BGS galaxy (right).

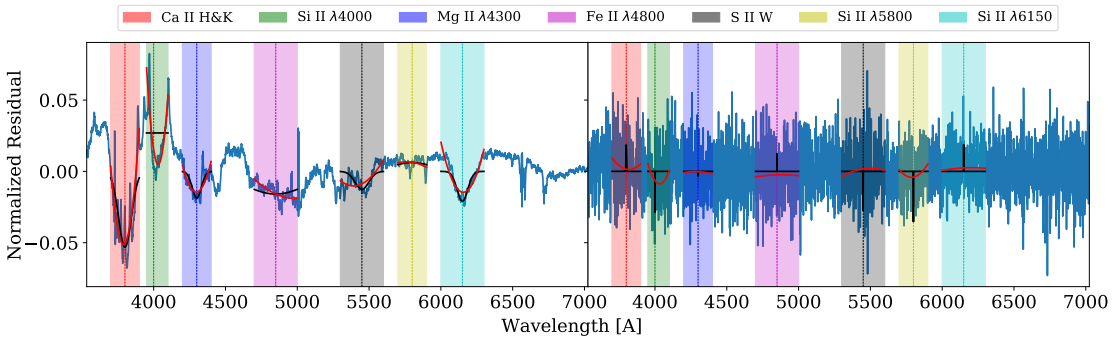

FIGURE 4.4: An example of the polynomial (red) and gaussian (black) fits to the feature ranges for a SN Ia host (left) and normal galaxy (right) residual spectrum.

The seven features, each corresponding to a different absorption feature in the SN Ia, are listed in Table 4.2. To define features in the residual, we define wavelength bands around the central wavelengths for each feature, which we define as the "feature range" in Table 4.2.

Within each feature range, we fit a polynomial and Gaussian to the residual flux to derive a measure of the amplitude of the absorption feature. The residual flux is normalized so that the amplitudes may be more appropriately compared across different spectra. The amplitude is defined as the difference between the maximum and minimum values of the fit in the feature range. An example for both a SN Ia host and a normal galaxy are shown in Figure 4.4. The seven amplitudes derived from these two fits are each used as a feature set in constructing a RF, which is trained to detect the difference in amplitude within these regions for SN Ia hosting and non-hosting galaxies. 75% of the 37,800 spectra (18,900 SN Ia hosts, 18,900 galaxies) are selected for training, and 25% for testing. We use 1,000 Monte Carlo cross-validations to obtain the final estimates of the RF's predictive ability.

This entire process is repeated for the bright, dim, faint, and very faint SN Ia host samples[1]. The results are displayed in Table 4.3. The accuracy of the RF is defined as the number of samples correctly identified divided by the total number of samples. The true-positive rate (TPR) is defined as the number of SN Ia hosts correctly predicted divided by the total number of SN Ia hosts. Likewise the true-negative rate (TNR) is defined as

[1]It is worth reemphasizing that all SN Ia spectra are at peak brightness.

SN Ia Sample	Accuracy [%]	TPR [%]	TNR [%]
Bright	$99.9941^{+0.0048}_{-0.0059}$	$99.9982^{+0.0018}_{-0.0018}$	$99.9900^{+0.0113}_{-0.0100}$
Dim	$99.9664^{+0.0194}_{-0.0229}$	$99.9736^{+0.0166}_{-0.0264}$	$99.9593^{+0.0235}_{-0.0200}$
Faint	$95.7503^{+0.1775}_{-0.1834}$	$94.2386^{+0.3227}_{-0.3175}$	$97.2613^{+0.2374}_{-0.2235}$
Very Faint	$56.9101^{+0.4913}_{-0.4956}$	$45.8725^{+0.7885}_{-0.7930}$	$67.9632^{+0.7778}_{-0.7345}$

SN Ia Sample	Accuracy [%]	TPR [%]	TNR [%]
Bright	$99.9621^{+0.0156}_{-0.0169}$	$99.9526^{+0.0368}_{-0.0265}$	$99.9716^{+0.0151}_{-0.0284}$
Dim	$98.5013^{+0.1091}_{-0.1137}$	$97.9972^{+0.1947}_{-0.2037}$	$99.0054^{+0.1387}_{-0.1508}$
Faint	$76.5863^{+0.3831}_{-0.3870}$	$64.7240^{+0.6662}_{-0.6565}$	$88.4572^{+0.5134}_{-0.5053}$
Very Faint	$43.9536^{+0.5365}_{-0.5227}$	$32.5798^{+0.8137}_{-0.8257}$	$55.3375^{+0.8972}_{-0.8769}$

TABLE 4.3: RF performance using amplitudes derived from polynomial (top table) and Gaussian (bottom table) fits in the residual spectra for the feature set.

the number of galaxies correctly classified divided by the total number of galaxies. The false-positive rate (FPR) and false-negative rate (FNR) are the complements of the TPR and TNR, respectively.

The RF does extremely well in both cases for the bright and dim samples, though the Gaussian fits perform significantly worse than the polynomial fits for the faint and very faint samples. The very faint (flux ratio = 0.01) is likely so dim that noise in the galaxy spectrum dominates the SN Ia features.

We find there are several disadvantage to this approach. First, there are several pre-processing steps needed to "condition" the data before the features are computed: an eigenbasis must be calculated, fit, and subtracted from the galaxy. The residual must be fit with functions that could be quite complex; the two cases we considered are relatively inexpensive compared to full template-fitting. Even so, the fits can be seen to be of poor quality, only marginally reflecting the actual shape of the features. Further, the eigenbasis does not represent a physical quantity - only the linear combination has a physical interpretation. The process of subtraction can distort or suppress the SN Ia features we are seeking to detect, increasing the difficulty and adding an additional complexity. Nonetheless, the RF is able to correctly classify SN Ia virtually every time in the bright and dim samples.

4.2.2 Features Using the Raw Spectrum

In order to decouple our analysis from relying on any fits at all, we consider features derived from the raw spectrum[2]. To do so, we derive the features from the pseudo-equivalent widths (pEWs) of the same features, which does not require any pre-processing of the data. For a spectrum with a well-defined continuum, the equivalent width (EW) is defined as the width of the spectral line such that the box extending from the x-axis to the continuum has an equal area to the area between the spectrum and the continuum. In our case, the spectral continuum is a non-trivial combination of continuum emission from the galaxy and continuum emission from the SN Ia, and is not readily available. Instead, we define a pseudo-continuum in place of the true continuum, and calculate the pEW by

$$\text{pEW} = \sum_{i=1}^{N} \left(1 - \frac{f_{\lambda_i}}{f_{c,\lambda_i}}\right) \Delta\lambda_i, \tag{4.5}$$

where f_{λ_i} is again the flux in wavelength bin λ_i, f_{c,λ_i} is the flux of the pseudo-continuum, and $\Delta\lambda_i$ is the spacing of the wavelength grid ($\Delta\lambda_i = 1$ Å uniformly over all our spectra). The sum is taken over the range of the pseudo-continuum.

We define the pseudo-continuum in the same manner as Nordin et al. (2011). The maximum flux in the lower and upper regions (see Table 4.2) for each feature is determined, and the straight-line connecting these two points is defined as the pseudo-continuum. Thus, the sum in Equation 4.5 is taken from the wavelength corresponding to the maximum in the lower region to that in the upper region. In order to avoid strong emission lines that happen to be in the lower or upper region dominating the maximum-finder and thus producing an inaccurate pseudo-continuum, a low-pass filter is applied to the spectrum. The maximum in the two regions is then calculated from this "smoothed" spectrum. An example of the pseudo-continuum and pEW for the 7 features is shown in Figure 4.5.

The pEW is then calculated, using the low-pass-filtered flux and the pseudo-continuum.

[2]Raw simply meaning no basis subtraction, etc.

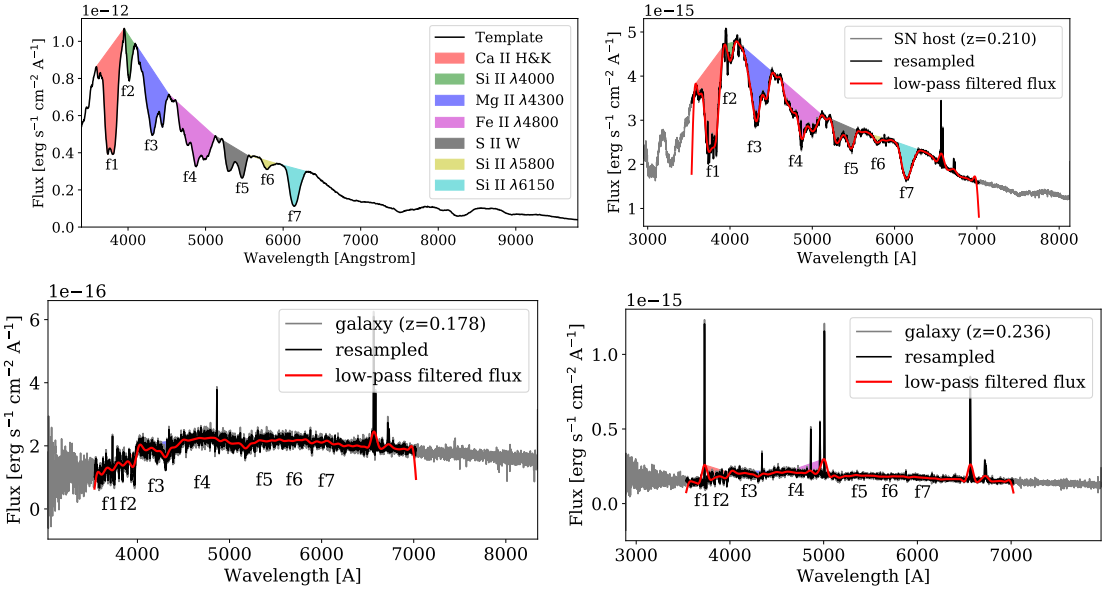

FIGURE 4.5: Definition of the pseudo-continuum (straight-lines) and pEWs (shaded areas) for the SN 2011fe template (top), bright SN Ia host (top center), regular galaxy (lower center) and galaxy with strong emission lines (lower). The full spectrum is shown in gray, and the resampled spectrum on the common wavelength grid is shown in black. The red line denotes the low-pass-filtered flux.

Figure 4.6 shows an example of the pEW calculations for the central two spectra in Figure 4.5. pEWs for a SN Ia host tend to be large, positive values, indicating strong absorption. pEWs for a galaxy tend to be very small (no emission/absorption) or occasionally negative (if an emission line is present in the feature).

Using the pEWs for the 7 features as the feature set, we train a RF to classify galaxies as hosting a SN Ia or not. The results are displayed in Table 4.4 for the bright, dim, faint, and very faint samples. The RF performs about as well or better than that with the polynomial amplitude feature set. Because the calculation of the pEWs is far simpler than the process for obtaining the line amplitudes, this approach seems to be the best candidate for finding SN Ia spectroscopically.

We also apply the RF to the luminosity distance-scaled sample of SN Ia to see how deep the RF can consistently detect SN Ia in a realistic scenario. In this case, the spectra are binned by redshift, such that each redshift bin contains 500 spectra. The RF is applied to the spectra in each redshift bin, and is cross-validated 100 times. The performance

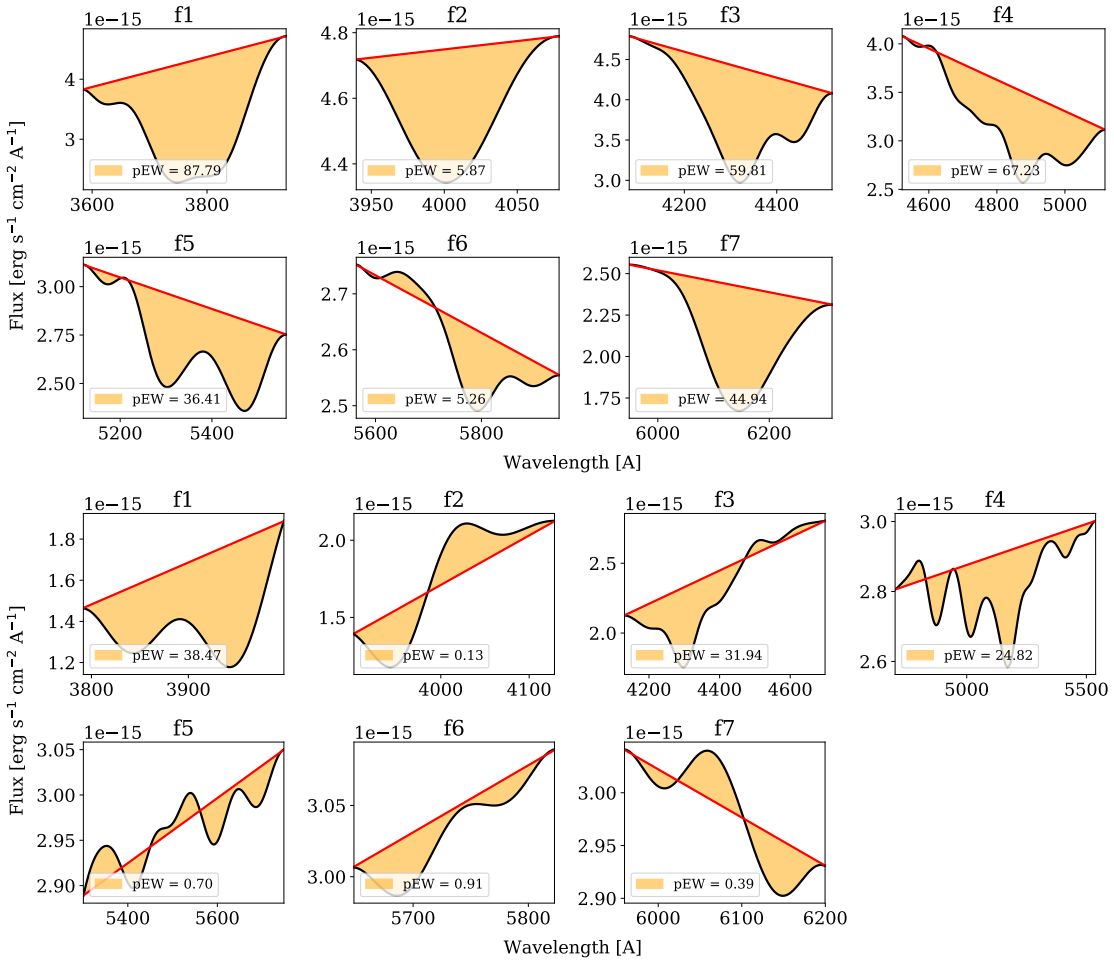

FIGURE 4.6: pEWs calculated from the feature set for an example bright SN Ia host (top) and a galaxy (bottom).

SN Ia Sample	Accuracy [%]	TPR [%]	TNR [%]
Bright	$99.9948^{+0.0055}_{-0.0052}$	$100.0000^{+0.0000}_{-0.0000}$	$99.9897^{+0.0111}_{-0.0103}$
Dim	$99.9619^{+0.0155}_{-0.0171}$	$99.9628^{+0.0265}_{-0.0372}$	$99.9610^{+0.0250}_{-0.0184}$
Faint	$94.8963^{+0.2009}_{-0.2142}$	$94.1917^{+0.3539}_{-0.3613}$	$95.6008^{+0.3068}_{-0.3034}$
Very Faint	$49.1067^{+0.4505}_{-0.4563}$	$40.7727^{+0.7558}_{-0.7316}$	$57.4514^{+0.7891}_{-0.7580}$

TABLE 4.4: RF performance using pEWs for the feature set.

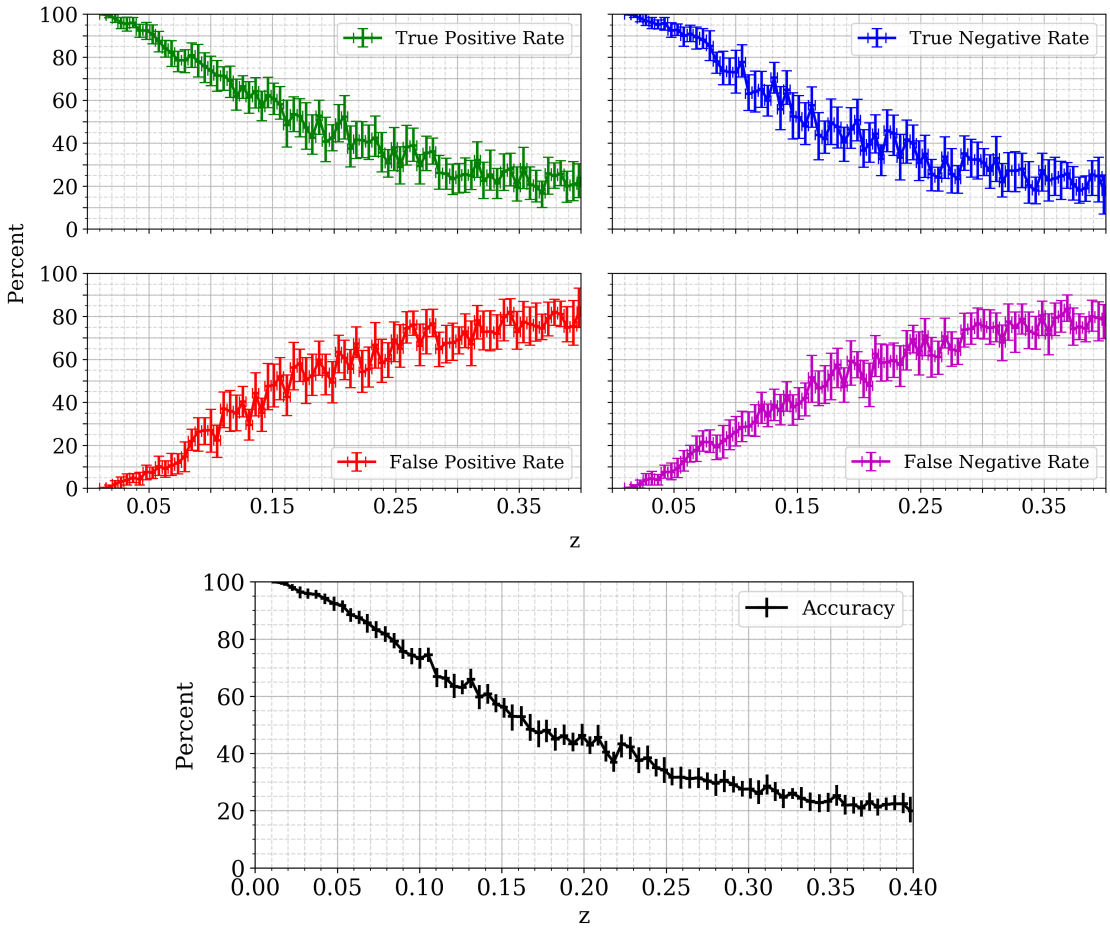

FIGURE 4.7: RF performance as a function of redshift, for realistic luminosity-distance scaled SN Ia spectra (all at peak brightness). The horizontal errorbars denote the bin-width. The vertical errorbars are the 16th and 84th percentiles.

statistics for each bin are plotted in Figure 4.7. We also bin with 1,000 and 5,000 spectra per bin, with 500 and 1,000 cross-validations respectively. We find no improvement in the quality of the predictions compared to 500 spectra per bin.

We see that out to about a redshift of $z \sim 0.1$ the RF has a $\geq 80\%$ TPR. This is about the extent of current SN Ia cosmology studies. Beyond this, a SN Ia becomes too faint relative to its host to be reasonably detected in this manner, consistent with the results from the flux ratio scaling.

We can also examine the relative importance of the individual pEW features. Figure 4.9 shows the relative importances across the bright, dim, faint, and very faint samples. From this it is clear that features f2 and f6 are relatively unimportant for bright SN Ia,

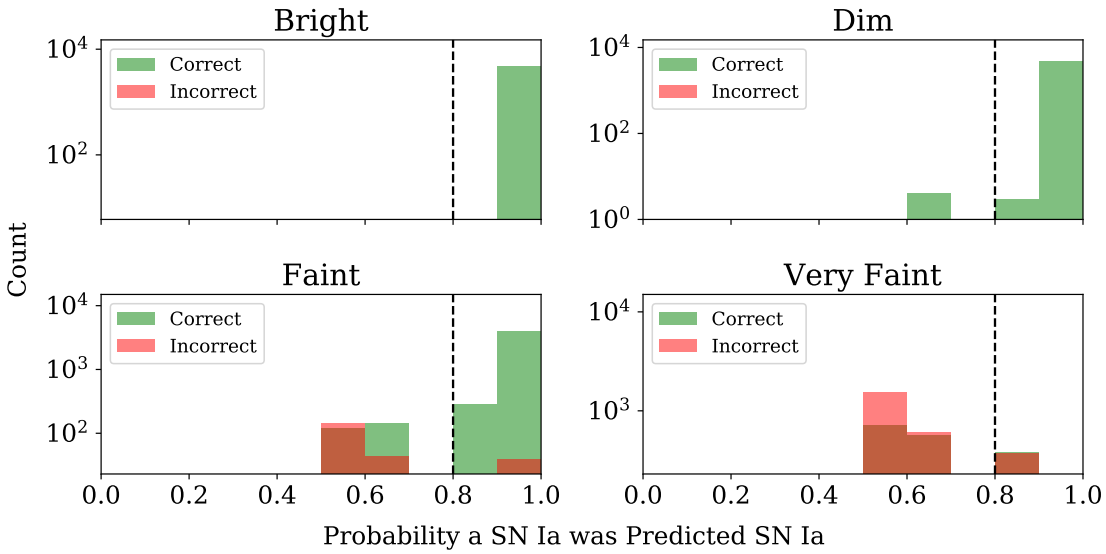

FIGURE 4.8: Distribution of predicted probabilities for correctly identified (green) and incorrectly classified (red) SN Ia hosts.

SN Ia Sample	TPR [%]	TPR after cut [%]	Rejected TP [%]
Bright	100.000	100.000	0.000
Dim	99.979	100.000	0.105
Faint	94.675	98.545	5.938
Very Faint	40.657	50.369	67.565

TABLE 4.5: True-positive rate (TPR) improvement after requiring a predicted probability of 0.8 or higher for a SN Ia candidate.

and that the strong Si II λ6150 feature (f7) used for SN type classification carries the most weight in this classification as well.

It is possible to improve the performance of the RF by enforcing a stricter prediction probability. Figure 4.8 shows the distribution of predicted probabilities for correctly and incorrectly predicted SN Ia hosts, for spectra that were SN Ia hosts. The improvement in the TPR obtaind by requiring a prediction probability of ≥ 0.8 is displayed in Table 4.5. This significantly improves the purity down to the faint sample with relatively few low-probability true-positive (TP) results being rejected. The very faint sample is again too faint to make a difference, but the cut does remove two thirds of very faint "claimed" SN Ia from the overall sample.

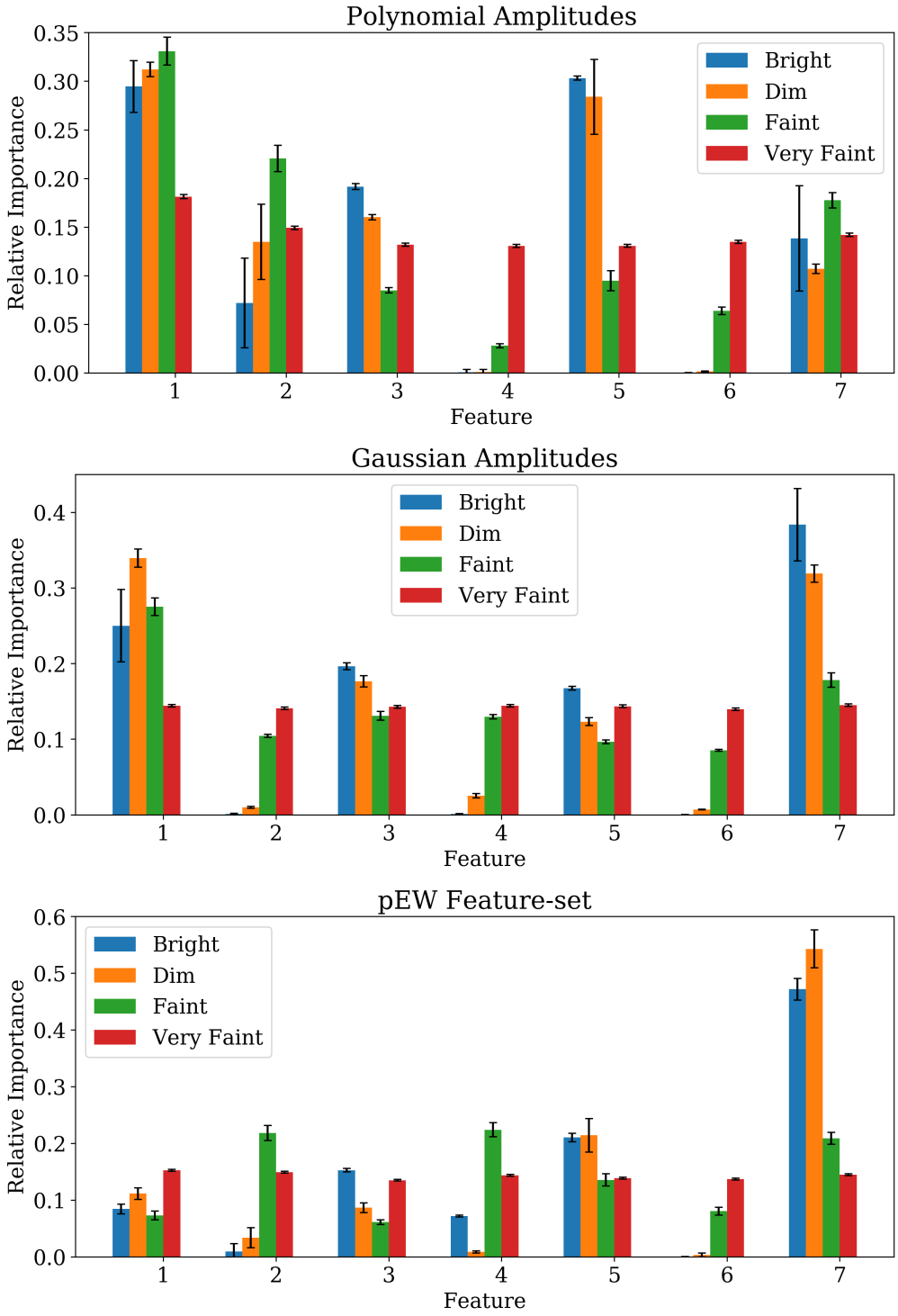

FIGURE 4.9: Relative importance of the 7 polynomial amplitudes (top), Gaussian amplitudes (center), and pEW features (bottom), for each of the flux ratio SN samples.

4.3 Outlier Detection Using Machine Learning

Rather than specifically look for one class of interesting spectra, one could instead look for any spectrum that is "unusual" in some regard. This is the concept behind outlier detection, which seeks to determine which items in a dataset are significantly different from a "normal" item in the dataset. A SN Ia in a galaxy spectrum is one example of an outlier spectrum, but many other physical processes could contribute to other types of outliers, for example type II supernovae, galaxy-galaxy gravitational lenses, galaxies with active galactic nuclei, or perhaps a new class of object altogether.

Random forests naturally lend themselves towards outlier detection applications. The basic idea behind detecting outliers is to define a form of distance measure for each item in the dataset that measures how similar it is to the general population. For a RF, this can be done under the context of unsupervised learning, that is without class labels known. The idea behind this is to define a second sample of data, called the *synthetic* data, that are drawn from a reference distribution derived from the observed data (which lack class labels), usually the marginal distributions for each feature (Shi and Horvath, 2006). In this way, the synthetic data will be distributed according to the same marginal distributions as the observed data, but feature-feature covariances are lost.

With the observed data labelled 'real' and the synthetic data labeled 'synthetic,' the task is now under the context of "supervised" learning. A RF is constructed by classifying 'real' and 'synthetic' data, essentially learning to tell apart data with correlated features from those without. Then, the real data are propagated through the trained forest, and each tree will predict either a *real* or *synthetic* label. Similar objects will follow similar paths through the trees in the forest. Thus, a similarity measure between each pair of objects can be assigned by counting the number of times the pair are both assigned a 'real' label by the same leaf of a given tree, normalized by the number of trees in the forest (Shi and Horvath, 2006). From this one could equivalently construct a distance measure by taking 1 - similarity. The distance between any object and itself is thus 0, and the distance between two objects that never reach the same leaf in any of the trees has a distance of 1.

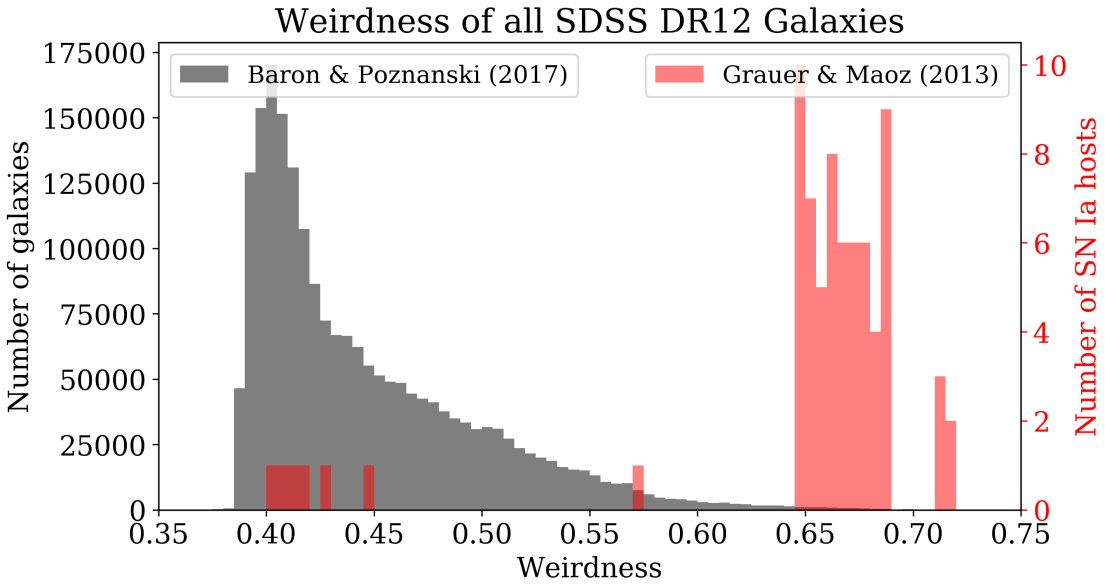

FIGURE 4.10: Weirdness distribution of SDSS galaxies from Baron and Poznanski (2017), with the weirdness scores of the SNe Ia discovered by Graur and Maoz (2013) shown in red.

This technique was used by Baron and Poznanski (2017) to find outliers in the SDSS data-release 12 (DR12) database. Using 2,355,926 median-normalized spectra, all interpolated to a common wavelength grid from 2250 Å to 10100 Å with 0.5 Å bins, Baron and Poznanski (2017) use all 15,700 flux values as features to construct their RF. They define a "weirdness" score by averaging the distance score for each galaxy across all of its pairs, resulting in a single number summarizing how "distant" the galaxy is from the usual. The weirdness distribution[3] they obtain from all two million SDSS galaxies is shown in Figure 4.10.

As it turns out, most of the SNe Ia discovered by Graur and Maoz (2013) have high weirdness scores assigned by Baron and Poznanski (2017) (Figure 4.10). Interestingly though, some do not, and this could potentially point to a population of false-positive detections or otherwise faint SNe. This could allow for a first cut in defining a sample of spectra to search for SNe Ia within. Coincidentally, Baron and Poznanski (2017) do not report on any of these SNe Ia, as they only consider the weirdest 400 galaxies, which happen to have a higher cutoff. Nonetheless, Baron and Poznanski (2017) find other

[3]The full weirdness table can be downloaded from http://wise-obs.tau.ac.il/cgi-bin/dovip/weird_galaxies.cgi.

interesting outliers such as galaxies with extreme emission lines, strong absorption lines, strange continua, double-peaked emission lines, galaxies with unusual gas kinematics, galaxies hosting supernovae, and more. With the goal of a generalized spectroscopic identification and classification pipeline, this represents a promising first step towards generalization.

Inspired by the success of this outlier detection, we attempted to train a RF using every flux value in the spectrum as a feature. The classification for SNe Ia hosts and non-hosts failed spectacularly, as the RF was no better than random guessing, even if the samples were the galaxies and bright-scaled SN 2011fe template (no injection!). We believe this is a result of the "curse of dimensionality," in that ~ 7000 features is too high relative to a sample size of only $37,800$. Perhaps a larger sample of galaxies could cure this, or perhaps while the full spectrum can be utilized for assigning distance measures and determining outliers, it is too-high dimensional to be used for traditional classification purposes.

Chapter 5

Conclusions and Future Work

In this thesis we have presented several methods for identifying SNe Ia in galaxy spectra. We constructed a simulated dataset using the DESI collaboration's simulation pipeline `desisim` to generate a mock catalog of galaxies and SN Ia hosting galaxies as seen by the DESI experiment. We considered an approach to detecting these SNe Ia using a principle component analysis to model the emission due to the galaxy, in which the model is subtracted, leaving a residual containing features resulting from phenomena not due to the galaxy. We defined feature-sets containing the amplitudes of absorption features due to SNe Ia, and used a random forest to classify galaxies as hosting a SNe Ia or not. We also considered a feature-set derived from the spectrum itself, without any modeling or subtraction, using the pseudo-equivalent widths of the same absorption features. We also briefly considered the application of RFs to outlier detection, in pursuit of more generalized spectroscopic classifications.

The RF trained with the pEWs performed the strongest and required the least pre-processing. We were able to accurately identify SNe Ia with ≥ 98 true-positive rate even down to the faint sample, which is built from SNe Ia scaled to a tenth of the host galaxies median flux. A flux difference of $1/10$ corresponds to a magnitude difference of 2.5. That is, Our model can accurately detect SNe Ia in galaxies up to 2.5 magnitudes brighter than the SN Ia itself. As can be seen in Figure 5.1, this corresponds to $\sim$40% of our simulated BGS sample. Hypothetically, some around half of the ELG sample and a small fraction of QSOs could also host detectable SNe Ia.

If we scale our BGS sample to the size of the equivalent survey DESI will obtain,

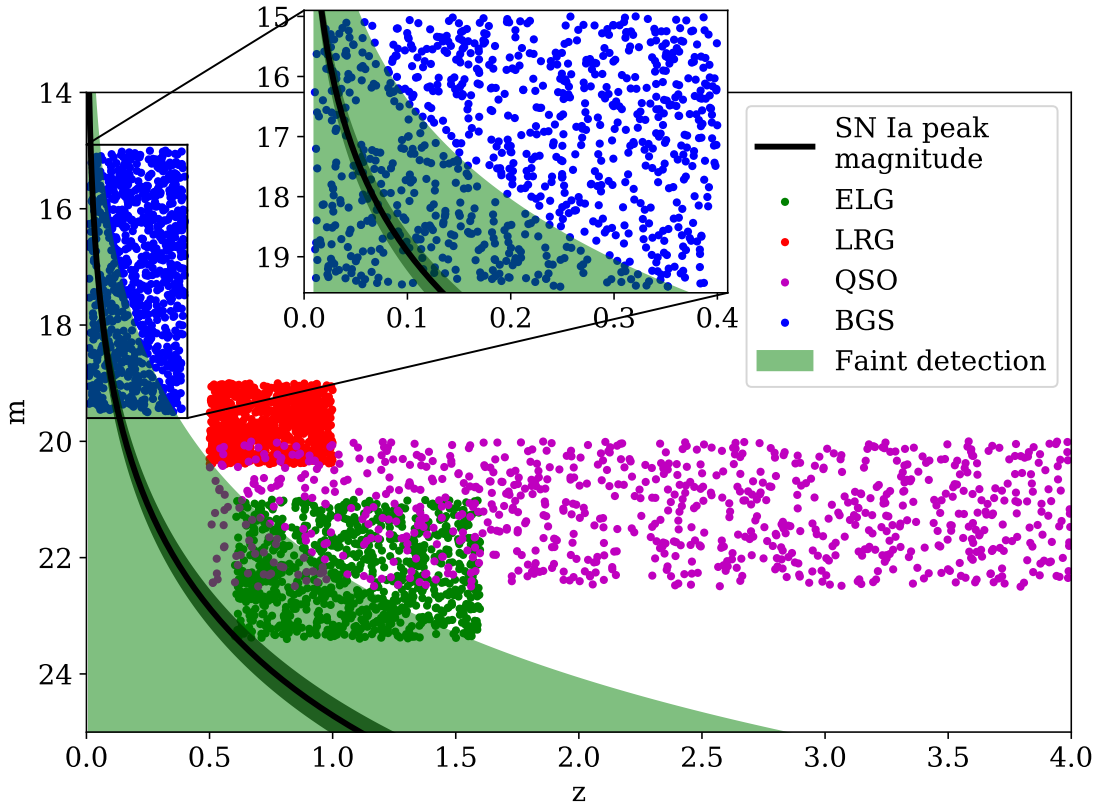

FIGURE 5.1: Apparent magnitude (m) vs. redshift (z, analogous to distance) for the four galaxy samples in the simulated dataset. The apparent magnitude of a SN Ia scaled to the luminosity distance at a given redshift is shown by the black curve, the band denoting the central 68% interval. The green envelope covers 2.5 magnitudes brighter (above) and infinitely dimmer (below) the SN Ia brightness curve. Any galaxy in the green envelope could host a SN Ia that is detectable by our method with $\geq 98\%$ true-positive rate.

consisting of 10 million galaxies, then assuming a SN rate of 1 per galaxy per century DESI is capable of detecting ∼10,000 SNe Ia over its five year survey. This is a substantial increase of 1-2 orders of magnitude over the current spectroscopic population of SNe Ia. Such a population will be essential for understanding the astrophysics of type Ia explosions themselves and for better understanding the dark energy powering the expansion of our universe.

5.1 Future Work

There are a number of improvements to this analysis to be considered in the future. Second order corrections to the quality of the spectra and SN Ia template injection will need to be considered, such as extinction/reddening and k-corrections (Hsiao et al., 2007). To improve the realism, the injection of the SN Ia template should be performed *before* folding the spectrum through the instrument response, rather than afterwards, to appropriately consider the Poisson errors in the flux. Because of the current state of `desisim`, this is not possible, but we expect this to be implemented in the near future.

A natural extension of this analysis is to test our identifier on SDSS archival data and confirm that we are able to re-discover supernovae found in past studies (e.g. Krughoff et al. (2011), Graur and Maoz (2013), Baron and Poznanski (2017)). While simulated data is attractive due to full *a priori* knowledge of the data, which allows for precise statistical evaluation of the algorithm (e.g. calculating the purity of a classified sample), it does not perfectly simulate the real thing. A question to consider will be training the RF on simulated data and testing on the real data, or training on real data using known SN sources from previous surveys to develop the classification scheme.

While there are many motivators towards detecting SNe Ia, other transients may too be spectroscopically identified and classified. Adding additional templates for other supernova classes, e.g. Ib/c, II, IIp, IIL, and including these in the simulated dataset will allow for measures of potential contamination. For example, how many type II supernovae are classified by the RF as type Ia. Currently, we are only considering two classes: Ia and not Ia.

With many large upcoming surveys like DESI, DES, and LSST, the increase in the quantity and complexity of astronomical data requires automated detection and classification pipelines. Improvement of the efficiency of the classifier could allow for implementation in realtime, which can provide the ability to detect new sources and alert other surveys for followup. Without the ability to detect in realtime, we are limited to archival searches and are thus limited in the amount of data we have, rather than the data we could take.

The ultimate step is to extending the algorithm to find other anomalous spectral features, such as gravitational lenses (Tsalmantza and Hogg, 2012) or weird AGN, similar to the method used by Baron and Poznanski (2017). To do so, we will need to think of how to find *not a galaxy*, rather than how to find *some specific transient*. A general and unassuming classifier will allow for the potential to make new discoveries we may have never expected. A successful algorithm will also identify bad spectra either due to instrumental issues or other errors, learn what those patterns look like, and avoid or even correct for them in the future.

Alternative approaches to the RF will need to be considered and evaluated, such as neural networks, clustering algorithms such as affinity propagation (Frey and Dueck, 2007; Mackenzie, Pichara, and Protopapas, 2016), and other machine learning outlier detection algorithms. Connections to photometric surveys should also be considered (du Buisson et al., 2015).

Lastly, this project has a natural extension involving public engagement. The citizen-science project Galaxy Zoo Supernovae demonstrated that members of the general public are remarkably good transient-spotters: 93% of supernovae were correctly identified by the public with no false-positives (Smith et al., 2011). These programs have proven track records of engagement and popularity with the public and offer a unique dual opportunity for public education and contribution to science. The data collected by citizen scientists can support actual results by providing confirmation or rejection of suspected outlier spectra. Thus, collaboration with these programs should be pursued and encouraged.

Appendix A

Example Datasets

In this appendix, five example spectra from each of the SN Ia samples are provided for reference. The left column contains five sample galaxies, the right the corresponding five SN Ia hosts. The center column shows the scaled SN Ia template that was injected to the galaxy to create the SN Ia host. All SN Ia are at peak brightness.

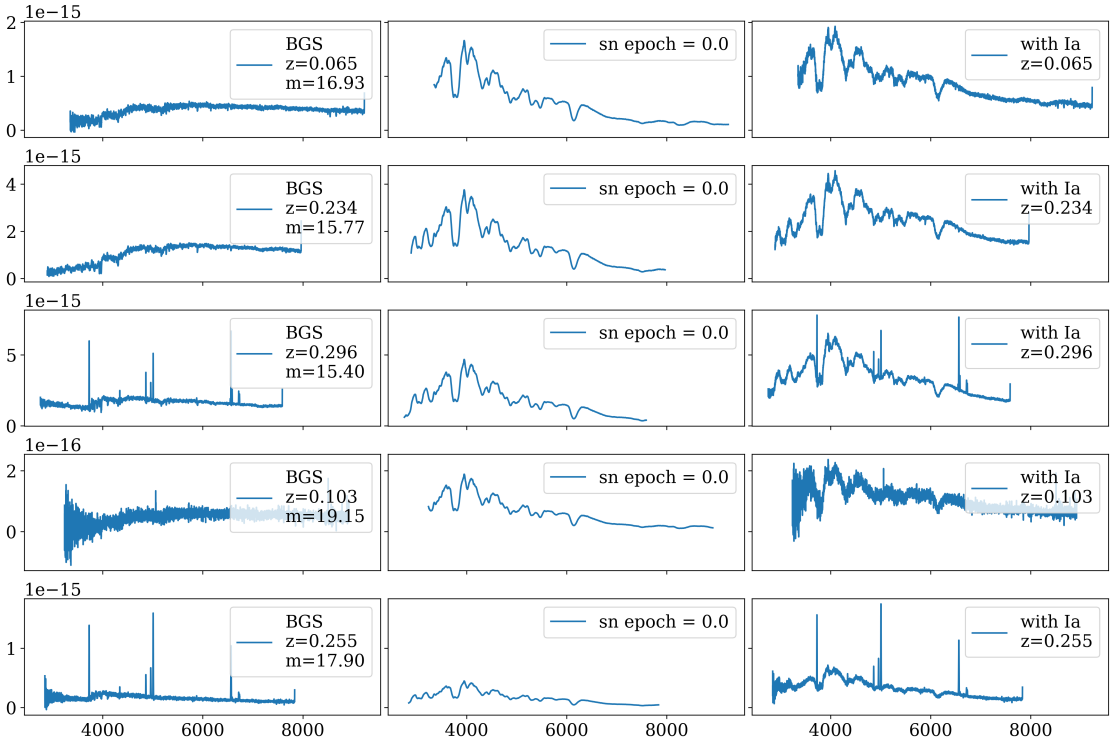

FIGURE A.1: Five spectra from the bright SN Ia sample.

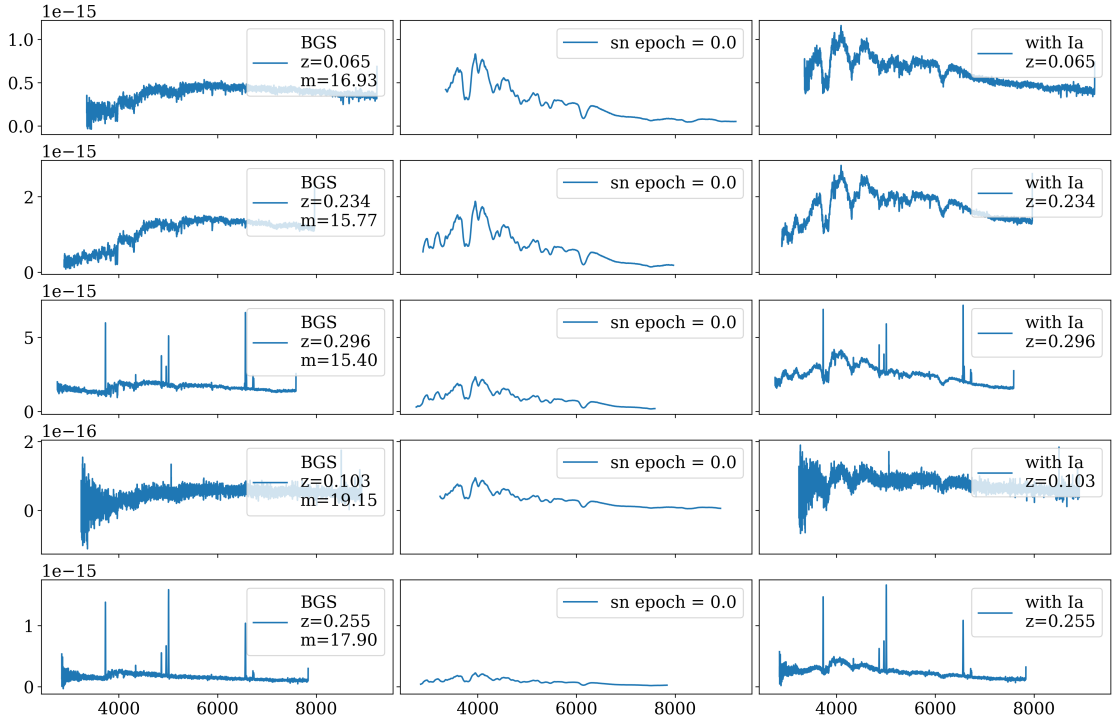

FIGURE A.2: Five spectra from the dim SN Ia sample.

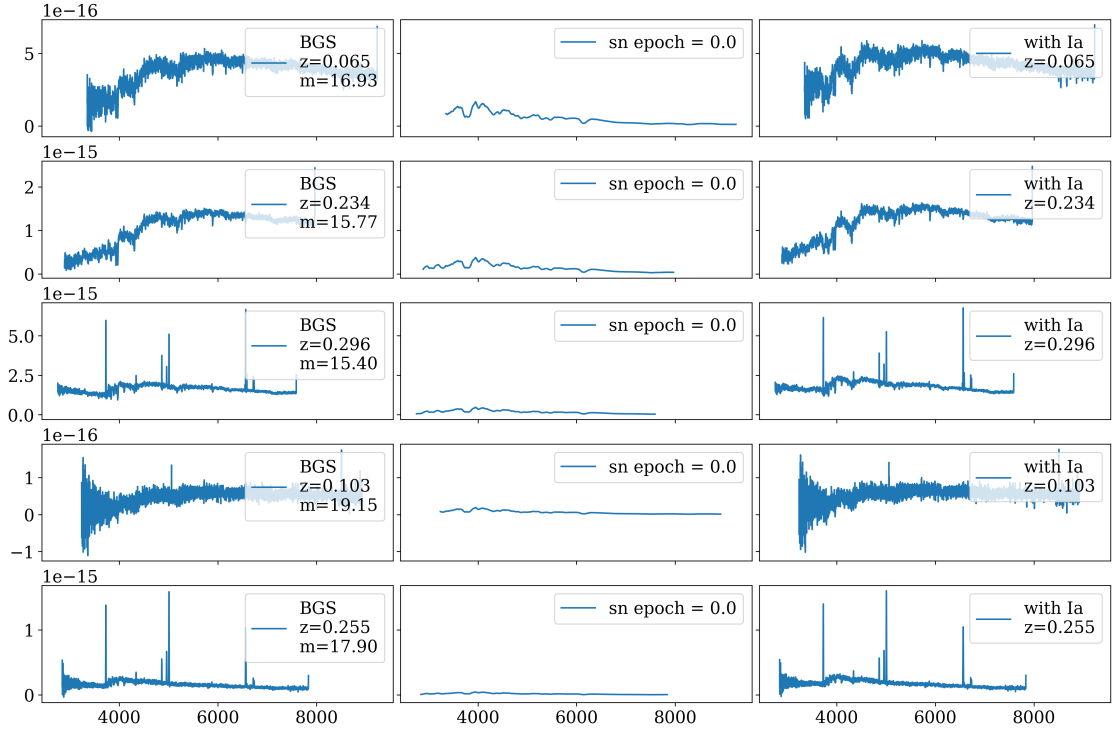

FIGURE A.3: Five spectra from the faint SN Ia sample.

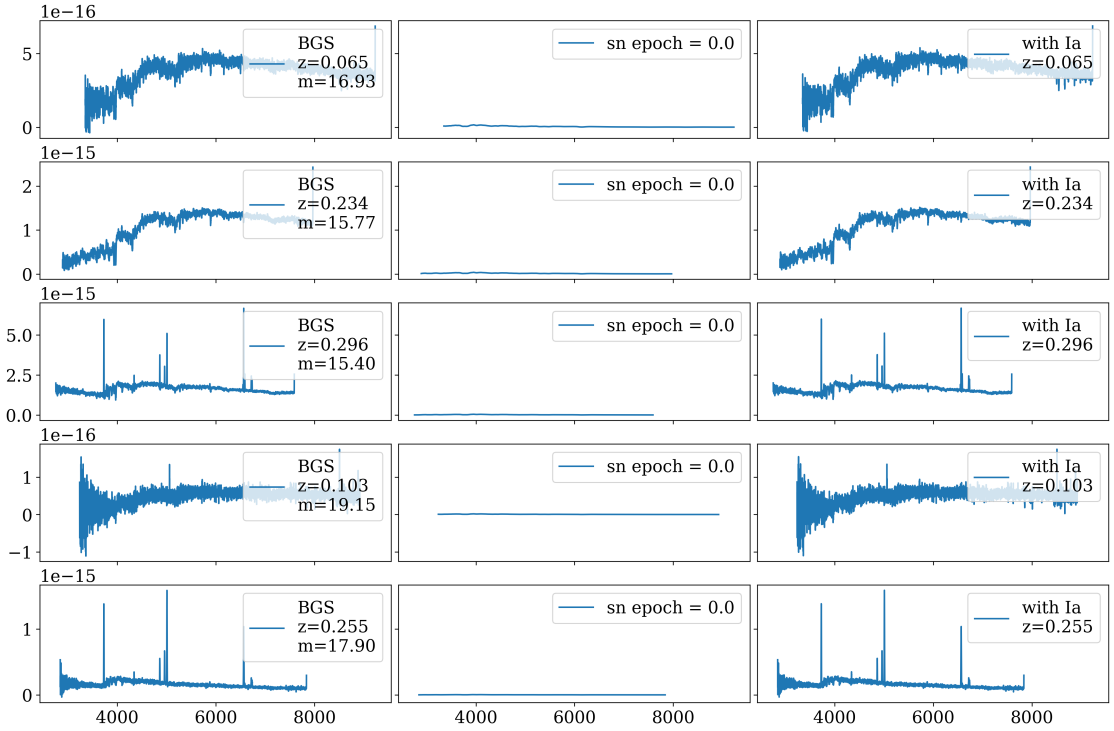

FIGURE A.4: Five spectra from the very faint SN Ia sample.

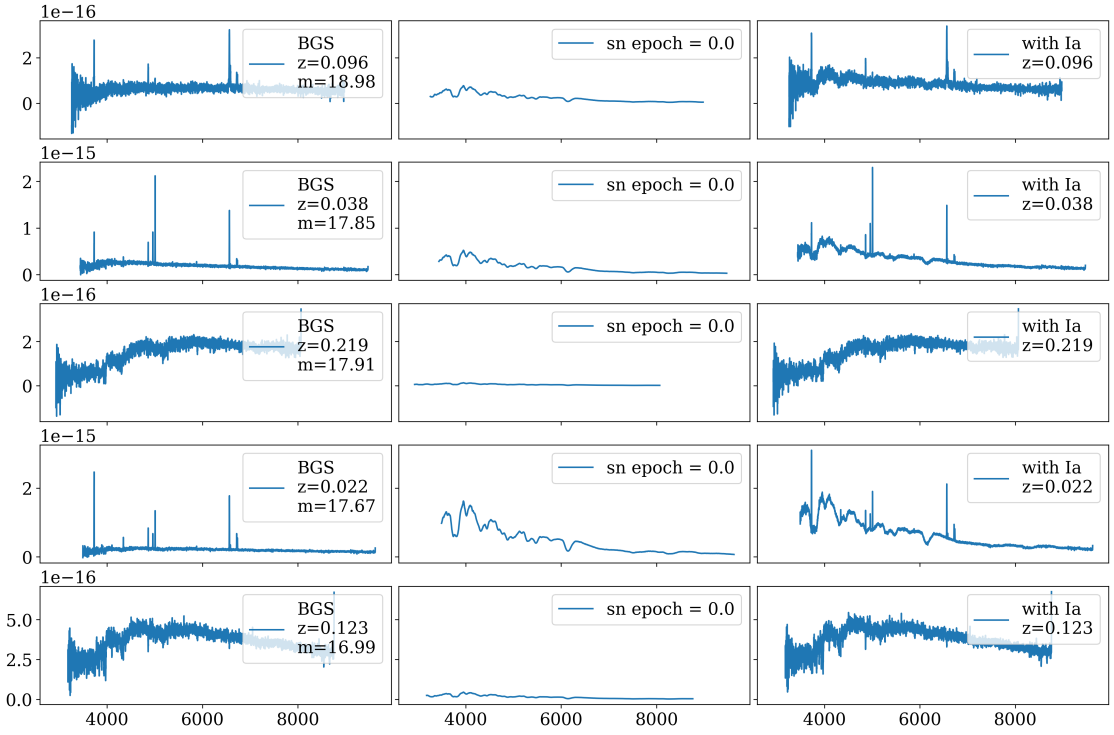

FIGURE A.5: Five spectra from the distance-scaled SN Ia sample.

Bibliography

Anderson, Wilhelm (1929). "Über die Grenzdichte der Materie und der Energie". In: *Zeitschrift für Physik* 56.11, pp. 851–856. ISSN: 0044-3328. DOI: 10.1007/BF01340146.

Badenes, C. and D. Maoz (2012). "The Merger Rate of Binary White Dwarfs in the Galactic Disk". In: ApJ 749, L11, p. L11. DOI: 10.1088/2041-8205/749/1/L11. arXiv: 1202.5472 [astro-ph.SR].

Baron, D. and D. Poznanski (2017). "The weirdest SDSS galaxies: results from an outlier detection algorithm". In: *Monthly Notices of the Royal Astronomical Society* 465, pp. 4530–4555. DOI: 10.1093/mnras/stw3021. arXiv: 1611.07526.

Branch, David et al. (1995). "In Search of the Progenitors of Type IA Supernovae". In: *Publications of the Astronomical Society of the Pacific* 107.717, p. 1019.

Breiman, Leo (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324.

Carroll, B.W. and D.A. Ostlie (2007). *An Introduction to Modern Astrophysics*. Pearson Addison-Wesley. ISBN: 9780805304022.

Chandrasekhar, S. (1931). "The Maximum Mass of Ideal White Dwarfs". In: ApJ 74, p. 81. DOI: 10.1086/143324.

Colgate, S. A. and C. McKee (1969). "Early Supernova Luminosity". In: ApJ 157, p. 623. DOI: 10.1086/150102.

DESI Collaboration et al. (2016a). "The DESI Experiment Part I: Science,Targeting, and Survey Design". In: *ArXiv e-prints*. arXiv: 1611.00036 [astro-ph.IM].

— (2016b). "The DESI Experiment Part II: Instrument Design". In: *ArXiv e-prints*. arXiv: 1611.00037 [astro-ph.IM].

du Buisson, L. et al. (2015). "Machine learning classification of SDSS transient survey images". In: *Monthly Notices of the Royal Astronomical Society* 454, pp. 2026–2038. DOI: 10.1093/mnras/stv2041. arXiv: 1407.4118 [astro-ph.IM].

Filippenko, Alexei V. (1997). "Optical spectra of supernovae". In: *Ann. Rev. Astron. Astrophys.* 35, pp. 309–355. DOI: 10.1146/annurev.astro.35.1.309.

Frey, Brendan J. and Delbert Dueck (2007). "Clustering by Passing Messages Between Data Points". In: *Science* 315.5814, pp. 972–976. ISSN: 0036-8075. DOI: 10.1126/science.1136800.

González Hernández, J. I. et al. (2012). "No surviving evolved companions of the progenitor of SN 1006". In: Nature 489, pp. 533–536. DOI: 10.1038/nature11447. arXiv: 1210.1948 [astro-ph.GA].

Graur, O. and D. Maoz (2013). "Discovery of 90 Type Ia supernovae among 700 000 Sloan spectra: the Type Ia supernova rate versus galaxy mass and star formation rate at redshift ~0.1". In: *Monthly Notices of the Royal Astronomy Society* 430, pp. 1746–1763. DOI: 10.1093/mnras/sts718. arXiv: 1209.0008.

Graur, O. et al. (2017). "LOSS Revisited. II. The Relative Rates of Different Types of Supernovae Vary between Low- and High-mass Galaxies". In: ApJ 837, 121, p. 121. DOI: 10.3847/1538-4357/aa5eb7. arXiv: 1609.02923 [astro-ph.HE].

Guillochon, J. et al. (2017). "An Open Catalog for Supernova Data". In: ApJ 835, 64, p. 64. DOI: 10.3847/1538-4357/835/1/64. arXiv: 1605.01054 [astro-ph.SR].

Hillebrandt, W. and J. C. Niemeyer (2000). "Type Ia Supernova Explosion Models". In: ARA&A 38, pp. 191–230. DOI: 10.1146/annurev.astro.38.1.191. eprint: astro-ph/0006305.

Howell, D. A. (2011). "Type Ia supernovae as stellar endpoints and cosmological tools". In: *Nature Communications* 2, 350, p. 350. DOI: 10.1038/ncomms1344. arXiv: 1011.0441 [astro-ph.CO].

Howell, D. A., M Sullivan, P. E. Nugent, et al. (2006). "The type Ia supernova SNLS-03D3bb from a super-Chandrasekhar-mass white dwarf star". In: *Nature* 443.7109, pp. 308–311.

Hsiao, E. Y. et al. (2007). "K-Corrections and Spectral Templates of Type Ia Supernovae". In: ApJ 663, pp. 1187–1200. DOI: 10.1086/518232. eprint: astro-ph/0703529.

Krughoff, K. S. et al. (2011). "Spectroscopic Determination of the Low-redshift Type Ia Supernova Rate from the Sloan Digital Sky Survey". In: *The Astrophysical Journal* 731, 42, p. 42. DOI: 10.1088/0004-637X/731/1/42. arXiv: 1102.1485.

Kushnir, D. et al. (2013). "Head-on Collisions of White Dwarfs in Triple Systems Could Explain Type Ia Supernovae". In: ApJ 778, L37, p. L37. DOI: 10.1088/2041-8205/778/2/L37. arXiv: 1303.1180 [astro-ph.HE].

Mackenzie, C., K. Pichara, and P. Protopapas (2016). "Clustering-based Feature Learning on Variable Stars". In: ApJ 820, 138, p. 138. DOI: 10.3847/0004-637X/820/2/138. arXiv: 1602.08977 [astro-ph.SR].

Maoz, D. and F. Mannucci (2012). "Type-Ia Supernova Rates and the Progenitor Problem: A Review". In: PASA 29, pp. 447–465. DOI: 10.1071/AS11052. arXiv: 1111.4492.

Nordin, J. et al. (2011). "Spectral properties of type Ia supernovae up to z ~ 0.3". In: A&A 526, A119, A119. DOI: 10.1051/0004-6361/201015705. arXiv: 1011.6227.

Parrent, J., B. Friesen, and M. Parthasarathy (2014). "A review of type Ia supernova spectra". In: Ap&SS 351, pp. 1–52. DOI: 10.1007/s10509-014-1830-1. arXiv: 1402.6337 [astro-ph.HE].

Planck Collaboration et al. (2016). "Planck 2015 results. XXIV. Cosmology from Sunyaev-Zeldovich cluster counts". In: A&A 594, A24, A24. DOI: 10.1051/0004-6361/201525833. arXiv: 1502.01597.

Riess, A. G. et al. (1998). "Observational evidence from supernovae for an accelerating universe and a cosmological constant". In: *Astron. J.* 116, pp. 1009–1038. DOI: 10.1086/300499. arXiv: astro-ph/9805201 [astro-ph].

Riess, A. G. et al. (2016). "A 2.4% Determination of the Local Value of the Hubble Constant". In: ApJ 826, 56, p. 56. DOI: 10.3847/0004-637X/826/1/56. arXiv: 1604.01424.

Schaefer, Bradley E. and Ashley Pagnotta (2012). "An absence of ex-companion stars in the type Ia supernova remnant SNR 0509?67.5". In: *Nature* 481, 164 EP –.

Shi, Tao and Steve Horvath (2006). "Unsupervised Learning With Random Forest Predictors". In: *Journal of Computational and Graphical Statistics* 15.1, pp. 118–138. DOI: 10.1198/106186006X94072.

Smith, A. M. et al. (2011). "Galaxy Zoo Supernovae". In: MNRAS 412, pp. 1309–1319. DOI: 10.1111/j.1365-2966.2010.17994.x. arXiv: 1011.2199 [astro-ph.IM].

Stoner, Edmund C. (1930). "LXXXVII. The equilibrium of dense stars". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 9.60, pp. 944–963. DOI: 10.1080/14786443008565066.

Timmes, F. X. and S. E. Woosley (1992). "The conductive propagation of nuclear flames. I - Degenerate C + O and O + NE + MG white dwarfs". In: ApJ 396, pp. 649–667. DOI: 10.1086/171746.

Truran, J. W., W. D. Arnett, and A. G. W. Cameron (1967). "NUCLEOSYNTHESIS IN SUPERNOVA SHOCK WAVES". In: *Canadian Journal of Physics* 45.7, pp. 2315–2332. DOI: 10.1139/p67-184.

Tsalmantza, P. and David W. Hogg (2012). "A Data-driven Model for Spectra: Finding Double Redshifts in the Sloan Digital Sky Survey". In: *The Astrophysical Journal* 753.2, p. 122.

Wang, B. and Z. Han (2012). "Progenitors of type Ia supernovae". In: New A Rev. 56, pp. 122–141. DOI: 10.1016/j.newar.2012.04.001. arXiv: 1204.1155 [astro-ph.SR].

Woosley, S. E., A. R. Kerstein, and A. J. Aspden (2011). "Flames in Type Ia Supernova: Deflagration-detonation Transition in the Oxygen-burning Flame". In: *The Astrophysical Journal* 734.1, p. 37.

Yip, C. W. et al. (2004). "Distributions of Galaxy Spectral Types in the Sloan Digital Sky Survey". In: *The Astronomical Journal* 128.2, p. 585.