The Astrophysical Journal Letters, 880:L22 (9pp), 2019 August 1 https://doi.org/10.3847/2041-8213/ab2edb

Optimal Classification and Outlier Detection for Stripped-envelope Core-collapse Supernovae

Marc Williamson[1], Maryam Modjaz[1,2], and Federica B. Bianco[2,3,4,5,6]
[1] New York University, New York, NY 10003, USA
[2] Center for Computational Astrophysics, Flatiron Institute, 162 5th Avenue, New York, NY 10010, USA
[3] Department of Physics and Astronomy, University of Delaware, Newark, DE 19716, USA
[4] Joseph R. Biden Jr. School for Public Policy and Administration, University of Delaware, Newark, DE 19716, USA
[5] Data Science Institute, University of Delaware, Newark, DE 19713, USA
[6] Center for Urban Science and Progress, New York University, USA

Received 2019 April 10; revised 2019 June 5; accepted 2019 June 9; published 2019 July 30

Abstract

In the current era of time-domain astronomy, it is increasingly important to have rigorous, data-driven models for classifying transients, including supernovae. We present the first application of principal component analysis to the photospheric spectra of stripped-envelope core-collapse supernovae. We use one of the largest compiled optical data sets of stripped-envelope supernovae, containing 160 SNe and 1551 spectra. We find that the first five principal components capture 79% of the variance of our spectral sample, which contains the main families of stripped supernovae: Ib, IIb, Ic, and broad-lined Ic. We develop a quantitative, data-driven classification method using a support vector machine, and explore stripped-envelope supernovae classification as a function of phase relative to *V*-band maximum light. Our classification method naturally identifies "transition" supernovae and supernovae with contested labels, which we discuss in detail. We find that the stripped-envelope supernovae types are most distinguishable in the later phase ranges of 10 ± 5 days and 15 ± 5 days relative to *V*-band maximum, and we discuss the implications of our findings for current and future surveys such as Zwicky Transient Factory and Large Synoptic Survey Telescope.

Key words: methods: data analysis – supernovae: general

Supporting material: data behind figure

1. Introduction

Supernova classification is a longstanding challenge in the astronomical community. The first spectral classification of supernovae (SNe) was introduced by Minkowski (1941), who defined two classes, Type I (hydrogen absent) versus Type II (hydrogen present). This broad criterion is still in use today, and multiple subclasses were added as the number of SNe spectra increased and spectral differences were observed (for a comprehensive review of SNe classification see Filippenko 1997; Gal-Yam 2017). In this work, we focus on stripped-envelope core-collapse supernovae (SESNe; Clocchiatti et al. 1997), which are the deaths of massive (>8 $M_\odot$) stars that have lost part or all of their outer hydrogen and helium layers. The diversity of the amount of these elements remaining in the outer envelopes of the stellar progenitors at the time of explosion is the likely explanation for the classification into three major SNe classes: Type Ib (spectra that have conspicuous He features), Type IIb (spectra showing strong H at early phases, He features at later phases), and Type Ic (no prominent H nor He features in spectra). For a more detailed review of SESNe see Filippenko et al. (1993), Matheson et al. (2001), Woosley et al. (2002), Modjaz et al. (2014), and Liu et al. (2016). Over the past 20 yr, the class of broad-lined SNe Ic (Ic-bl) has emerged with members showing spectra devoid of strong lines of H and He, but with broad lines that indicate expansion velocities between 15,000 and 20,000 km s^{-1} (Modjaz et al. 2016; Prentice & Mazzali 2017; Sahu et al. 2018). In addition, the Ic-bl type is the only SN type associated with long-duration gamma-ray bursts (for reviews see Woosley & Bloom 2006; Modjaz 2011; Cano et al. 2017).

Current SESNe classification methods can be grouped broadly into two categories: template matching and specific feature techniques. The most used template-matching algorithms are the Supernova Identification code (SNID; Blondin & Tonry 2007) and Superfit (Howell et al. 2005). These codes match new spectra to a library of previously classified supernovae using cross-correlation and chi-squared statistics, respectively, yielding a quantitative measure of similarity between spectra of previously known SNe and the spectrum of a new transient. By incorporating more than just the best match into a classification scheme (e.g., Quimby et al. 2018), template matching can distinguish the major SESNe classes. However, template matching has some downsides. It is difficult to gain physical insight into stellar progenitors from a simple similarity measure. In addition, template-matching classification methods do not directly yield a physical understanding of the differences between different classes. The second category of classification techniques focuses on characterizing specific spectral features (i.e., line depth or width and line intensity or velocity) at particular wavelengths (Prentice & Mazzali 2017; Sun & Gal-Yam 2017). These specific feature techniques allow for more physical interpretation than template matching, but they do not use all of the information available in a spectrum.

In this paper, we propose a new classification technique for SESNe using principal component analysis (PCA) combined with support vector machine (SVM). PCA is a dimensionality reduction algorithm that linearly transforms data in order to capture as much information as possible in the smallest number of transformed features, called principal components (PCs). PCA has been previously applied to attempt to understand the diversity of SNe Ia subtypes (Cormier & Davis 2011; Sasdelli et al. 2014) and nebular phase superluminous supernovae (Nicholl et al. 2019), but this is the first application of PCA to SESNe in the photospheric phase. After applying a PCA decomposition to our SESNe spectral data set, we use a multiclass linear SVM, a supervised learning method, to classify our SNe. This work is the first application of such machine-learning techniques to spectroscopically classifying SESNe.

Our PCA and SVM based algorithm allows continuous, quantitative classification that reflects the physical properties of SESNe stellar progenitors. Instead of the traditional SN classification with four discreet classes (IIb, Ib, Ic, Ic-bl), our classification method facilitates better understanding of which SNe are representative of their class, and which are "transition" objects, and comparison between the SESNe mean spectra and our constructed eigenspectra allows us to physically interpret our results. New and upcoming data releases by the Berkeley group (Shivvers et al. 2018) and the Palomar Transient Factory (PTF; Fremling et al. 2018; Taddia et al. 2018), and new transient observing projects like the Zwicky Transient Factory (ZTF; Bellm et al. 2019) and the Large Synoptic Survey Telescope (LSST; Ivezic et al. 2008) will drastically increase the number of SESNe spectra. In this new data-rich context, a continuous and data-driven classifier will be crucial for addressing some of the most interesting outstanding questions pertaining to SESNe.

2. SESNe Spectral Data Set

In this section, we describe the spectral data set used in this work and the preprocessing applied to the data before our analysis is performed. We expand the SESNe spectral library produced and compiled in Modjaz et al. (2014), Liu & Modjaz (2014), Liu et al. (2016), and Modjaz et al. (2016) to include available spectra from SNe published through 2018 August. We use the same criteria for inclusion of new data as Liu et al. (2016): well-typed SNe with light curves for which a date of maximum can be extracted. The data set contains 160 SNe and 1551 spectra. We exclude SNe Ib-n, SNe Ib-Ca, superluminous supernovae, and SNe that transition between normal and excluded types. We restrict the spectra in our sample to the optical wavelength range 4000–7000 Å since the vast majority of our SNe have observed fluxes in this wavelength range, and this range contains features of both H and He that drive the classification. For newly added SN spectra obtained from the literature or directly from authors, we follow the same preprocessing steps detailed in Liu et al. (2016) that were used in subsequent papers of our group (Liu et al. 2016; Modjaz et al. 2016). The preprocessing is briefly summarized as follows: when newly added spectra lack a date of V-band maximum (but do have a date of maximum in other bands), we convert their date of maximum to the V band using the process described by Bianco et al. (2014). Spectra are redshift corrected when necessary, and the continuum removal and normalization (spectra are scaled by their means to have relative fluxes) is performed with tools within the SNID framework (Blondin & Tonry 2007). In the few cases where telluric lines are present in the spectra, the tellurics are removed using linear interpolation consistent with the procedure in Liu et al. (2016). Small gaps in the spectra are similarly interpolated before a fourier-based smoothing is applied (Liu et al. 2016). The bandpass filter used by SNID for classification purposes is not applied. A summary of our data set can be found in Table 1, and the SNID templates of the newly added SNe are released on our GitHub page.[7]

3. Methods

In this section, we present a brief background on the two machine-learning methods used in our analysis, PCA and SVM, as well as details on our specific application. For both methods, we use the `scikit-learn`[8] implementation (Pedregosa et al. 2011). For a detailed review of PCA theory see Pearson (1901) and Jolliffe (2011), and for a detailed review of SVM theory see Vapnik (1998). Our research is reproducible: all code and raw data are accessible on GitHub.[9]

3.1. PCA—Derivation of Eigenspectra

PCA is a dimensionality reduction technique based on singular value decomposition of a data matrix. The principal components are the eigenvectors of the covariance matrix of the data, and are therefore orthonormal. Each PC is a linear combination of the original data features (normalized fluxes) and therefore has the same wavelength range as our original data. We therefore use the term "eigenspectra" to describe the PCs and discuss their physical interpretation in Section 4. The eigenspectra are ordered according to how much variance from the mean of the data set each component captures. Thus, the original spectra can be projected onto a subset of the eigenspectra while maximizing the amount of information retained. Figure 1 shows the cumulative amount of variance of the entire data set captured as a function of the number of PCs. The first five eigenspectra contain 79% of the variance. Figure 2 shows an example supernova, SN2011ei (type IIb), reconstructed using increasingly larger numbers of eigenspectra. In the top panel, only the first five eigenspectra are used, and the large-scale spectral features are almost entirely reconstructed. For the purpose of classification, we mostly care about the large-scale features, so considering only the first five eigenspectra of our PCA decomposition is a good first step to reduce the complexity of the problem.

Since SNe change over time, in this work we apply a PCA decomposition to four different phase ranges of spectra: 0 ± 5, 5 ± 5, 10 ± 5, and 15 ± 5 days relative to the V-band maximum. We present and discuss in detail the eigenspectra for the phase range $t_{V_{\max}} = 15 \pm 5$ days in Section 4.1. The time dependence of the eigenspectra as a function of phase is discussed in Section 4.2, but in general we find that there is very little change in the large-scale features of a given eigenspectrum over time.

3.2. SVM—A New Approach to SESNe Classification

For each of the four phase ranges in this work, we train a multiclass linear SVM without class weights and using the L2 (i.e., "squared hinge") loss function on the 2D projection of SESNe spectra onto each pair of the first 10 eigenspectra in order to understand which eigenspectra are most useful for classification. Specifically, we use the LinearSVC class from

[7] https://github.com/nyusngroup/SESNtemple/tree/master/SNIDtemplates
[8] https://scikit-learn.org/stable/index.html
[9] https://github.com/nyusngroup/SESNspectraPCA

Table 1
SESNe Data Set (Includes Compilations from L&M14,[a] L16,[a] M16,[a] and New Additions Below)

Phase[a]	Ib (N_{SNe}, N_{Spec})	IIb (N_{SNe}, N_{Spec})	Ic (N_{SNe}, N_{Spec})	Ic-bl (N_{SNe}, N_{Spec})
0 ± 5	(28, 81)	(21, 62)	(27, 79)	(17, 74)
5 ± 5	(22, 68)	(19, 54)	(21, 61)	(17, 59)
10 ± 5	(23, 54)	(18, 41)	(21, 47)	(15, 36)
15 ± 5	(19, 44)	(17, 35)	(18, 40)	(13, 30)

New SNe (Added to Liu et al. Sample)

SN Name	SN Type	Phases[a]	References
SN2010as	IIb	−14, −13, −12, −11, −10, −9, −9, −9, −9, −8, −8, −8, −6, −6, −6, −5, −5, −5, 6, 19, (+6)	F14
SN2011hs	IIb	−9, −8, −8, −7, −6, −5, −5, −5, 4, 5, 8, 25, 25, 26, 27, 28, 28, 53, 58, (+4)	B14
SN2012au[b]	Ib	−6, −1, 10, 21, 33, 48, 57, 67, 73, (+2)	T13
SN2012P	IIb	−11, −8, −7, −2, 1, 8, 26, 29, 31, (+1)	F12
SN2013df	IIb	−14, −11, −4, −4, −4, 0, 4, 11	S16,C16
SN2013ge	Ic	−15, −14, −13, −4, −3, 5, 6, 12, 13, 15, 16, 19, 30, 33, 34, 37, 38, 40, 43, 44, 46, 46, 62, 65, 67, 70, 70, 71, (+9)	D16
SN2014ad	Ic-bl	27, 27, 37, 37	YG12
LSQ14efd	Ic	−12, −11, −4, −4, −4, 4, 4, 17, 17, 17, 23, 23, 23, 32, 32, 32, 32	S15
iPTF15dtg	Ic	−16, −2, 6, 15, 45, 64, 78, (+1)	T16
SN2016coi	Ic-bl	−13, −10, −8, −2, −1, 0, 2, 6, 7, 10, 23, 24, 30, 42, 62, 85, (+3)	P18
SN2016gkg	IIb	−18, −18, −18, −17, −16, −15, −14, −11, −10, −7, −5, −1, −1, 0, 1, 15	T17
SN2017ein	Ic	−7, 12, 15, 18, 22, 38, 47, 51, 53	VD18

Notes.
[a] Phases are rounded to the nearest integer and are in rest frame, relative to the date of *V*-band maximum. The number of nebular phase spectra (phase >90 days) are included in parentheses, but they are not used in our analysis.
[b] The spectra of SN2012au could not be included in our PCA analysis because they were too noisy.

References. L&M14—Liu & Modjaz (2014), L16—Liu et al. (2016), M16—Modjaz et al. (2016), F14—Folatelli et al. (2014), B14—Bufano et al. (2014), T13—Takaki et al. (2013), F12—Fremling et al. (2016), S16—Szalai et al. (2016), C16—Childress et al. (2016), D16—Drout et al. (2016), YG12—Yaron & Gal-Yam (2012), S15—Smartt et al. (2015), T16—Taddia et al. (2016), P18—Prentice et al. (2018), T17—Tartaglia et al. (2017), VD18—Van Dyk et al. (2018).

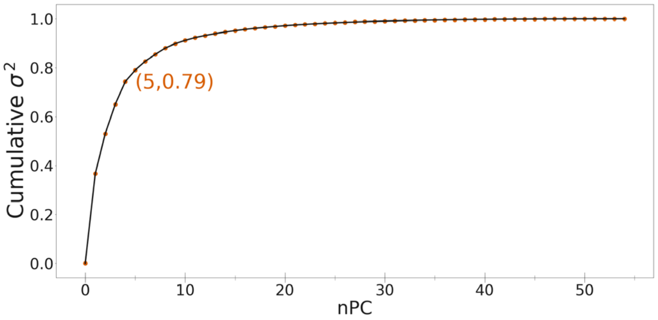

Figure 1. Cumulative fraction of variance of the entire SESNe data set captured by nPC eigenspectra. The first five eigenspectra capture 79% of the sample variance.

`scikit-learn`, which implements SVM classification using LIBLINEAR (Fan et al. 2008) and employs the "one-versus-rest" approach to multiclass labeling and the "winner-take-all" approach to multiclass predictions: a binary linear SVM is trained to distinguish each class of SESNe from the rest of the population, and these binary classifiers are combined to make final decisions on predicting the labels of new data. Each binary SVM determines the optimal hyperplane that separates one class from the rest of the data. For each 2D projection, we randomly generate multiple train-test splits of the data (a random subset of 70% of the data is used to train the SVM, while the remaining 30% is used to test the ability of the SVM to accurately predict SNe classes). Using multiple train-test splits on each 2D projection allows us to report a mean test score for the SVM and to gain insight into the uncertainty of

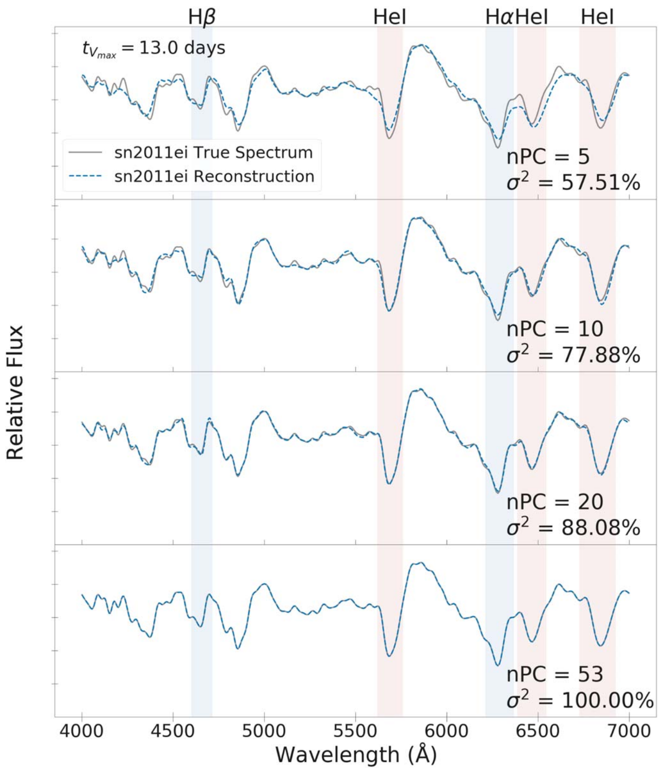

Figure 2. Reconstructions of the spectrum of SN type IIb SN2011ei (Milisavljevic et al. 2013) at phase $t_{V_{max}} = 13$ days. An increasing number of eigenspectra (nPC) is used to reconstruct the original spectrum from top to bottom. As nPC increases, more features are captured, but five eigenspectra already capture the H and He features (indicated by shaded regions).

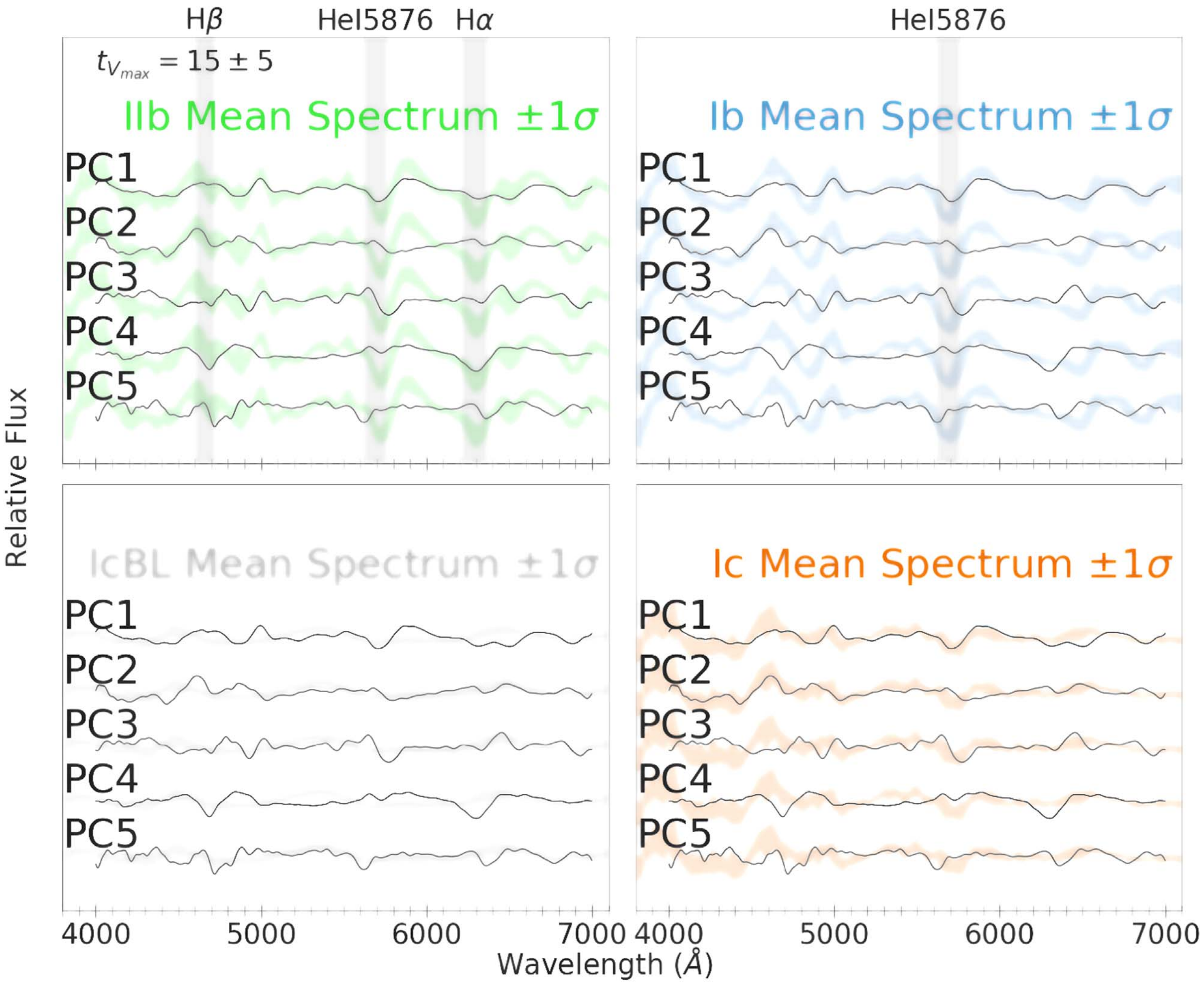

Figure 3. Comparison of the first five eigenspectra at phase $t_{V_{\max}} = 15 \pm 5$ days, constructed using data of all SESNe types, with the mean spectra (Liu et al. 2016; Modjaz et al. 2016) for types SNe IIb (upper left), Ib (upper right), Ic-bl (lower left), and Ic (lower right). The eigenspectra are scaled by a factor of 2 and sign choice is made to facilitate comparing the relative structure of the principal components vs. the mean spectra. PC1 and PC3 have a strong trough that lines up with the He I5876 absorption feature in types IIb and Ib SNe. PC4 has strong troughs that line up with the Hα and Hβ absorption features in the type IIb mean spectrum.

the SVM linear decision boundaries. The results of our SVM classification are discussed in Section 5.

4. Physical Interpretations of Eigenspectra

One of the major benefits of our PCA and SVM based classification method is that we can physically interpret the eigenspectra using mean spectra of each of the SESNe classes. This allows us to understand why the SVM identifies certain eigenspectra as better classifiers than others, and how this behavior changes as a function of phase.

4.1. Comparing Eigenspectra to SESNe Mean Spectra

The first few eigenspectra are the most important building blocks for reconstructing a spectrum from our data set. Therefore, in order to understand any strong eigenspectra features, we compare the first five eigenspectra for the phase range $t_{V_{\max}} = 15 \pm 5$ days to the mean spectra for each of the four SESNe types, presented in Liu et al. (2016) and Modjaz et al. (2016). The first five eigenspectra are plotted in Figure 3, along with the mean spectra for the four SESN subtypes. The principal components are naturally normalized, and we choose the sign of each component to properly represent the absorption features they capture. We highlight a few important features of each of the first five eigenspectra:

1. PC1 has a strong trough that lines up with the He I5876 absorption feature present in both type IIb and Ib mean spectra, as well as the absorption feature in the Ic mean spectrum (the cause of which is debated; Dessart & Hillier 2010).
2. PC2 matches the Ic mean spectra closely up to $\lambda \approx 5500$ Å.
3. PC3 has small troughs in the Hα and Hβ regions in addition to a stronger trough in the He I5876 region.
4. PC4 has strong troughs in the Hα and Hβ regions, but lacks a strong He I5876 feature.
5. Due to the broadening of their features, SNe Ic-bl are effectively nearly featureless spectra, which result in a much smoother average spectrum than any of the first five PCs.

These similarities between the eigenspectra and the SESNe mean spectra provide an excellent context to interpret the SVM

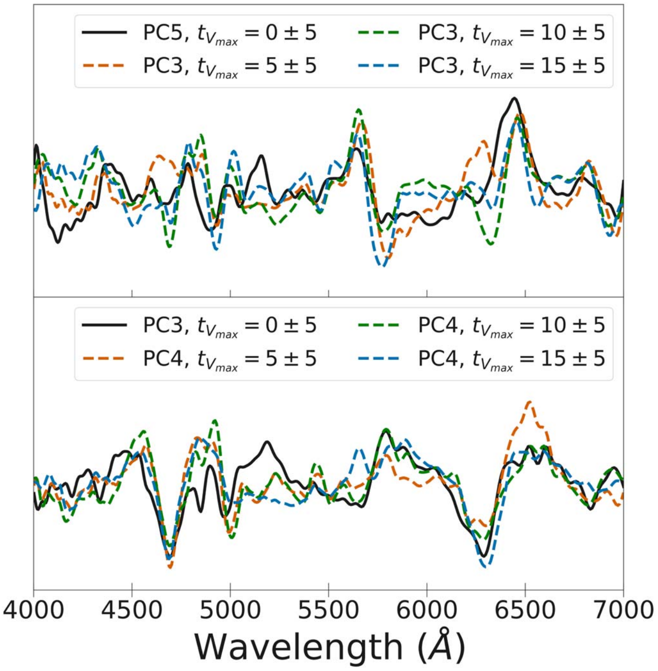

Figure 4. Change in eigenspectrum order between $t_{V_{\max}} = 0 \pm 5$ days vs. later phase ranges. PC5 at early times is equivalent to PC3 at later times as they capture the same features: Hα, Hβ, and He I5876 as discussed in Section 4.1. Similarly, PC3 at early times is equivalent to PC4 at later times because they both primarily capture Hα and Hβ absorption. Otherwise, the important large-scale features do not change with time.

classification. From Figure 3, we see that all of the SESNe types except Ic-bl have an absorption feature near $\lambda \approx 5876$ Å (although this feature is most likely not due to helium for the Ic type). Moreover, as shown in Liu et al. (2016), this feature exists in the IIb, Ib, and Ic mean spectra even at early phases. Therefore we conclude that PCA generates eigenspectra that match previously identified important SESNe spectral features.

4.2. Time Evolution of Eigenspectra

In Section 4.1 we present the eigenspectra only for the phase range $t_{V_{\max}} = 15 \pm 5$ days because we find the SESNe types to be maximally separated at this phase, as we show in Section 5. Here we discuss how the eigenspectra change as a function of time. We have calculated and compared the first five eigenspectra for each of the phase ranges 0 ± 5 days, 5 ± 5 days, 10 ± 5 days, and 15 ± 5 days, relative to the V-band date of max. We find that there is very little change for a given eigenspectrum across different phases. However, there is a slight change in the ordering of the first five eigenspectra between the later phase ranges and the $t_{V_{\max}} = 0 \pm 5$ day phase range. Figure 4 shows that PC5 at phase $t_{V_{\max}} = 0 \pm 5$ days corresponds (i.e., is most similar) to PC3 of the later phase ranges, and PC3 at phase $t_{V_{\max}} = 0 \pm 5$ corresponds to PC4 of the later phases. In the later phase ranges, PC3 is the eigenspectrum with weak troughs in the Hα and Hβ regions and a strong trough in the He I5876 region. Thus, it is not surprising that this eigenspectrum captures less variance of the sample in the earliest phase range. Liu et al. (2016) showed that the pseudo-equivalent line width (pEW) of He I5876 in SNe types IIb and Ib are at their lowest values near the V-band maximum and increase over time. PC4 in the later phases, which consists of two strong troughs at the Hα and Hβ wavelengths, is more highly ranked in the $t_{V_{\max}} = 0 \pm 5$ phase range because the Hα absorption feature is very strong in type IIb spectra even at early phases.

5. SVM Classification Results

Our goal is to create a method that reproduces the standard empirical classification scheme that classifies SNe spectra using the H and He features. We apply SVM to every 2D projection of the first 10 eigenspectra (following the procedure in Bianco et al. 2016), and we find that the highest test scores (see Section 3.2) are always associated with a pair of the top five eigenspectra. These results are consistent with both Figures 1 and 2, where the first five eigenspectra are sufficient to capture 79% of the spectral variance in our sample and reproduce the spectrum of SN2011ei, respectively. In Figure 5, each panel corresponds to a different phase range, and in each panel we show the 2D plane that leads to the highest classification score. In the case where one phase range has multiple optimal planes (i.e., test scores are consistent within 1σ), we choose the eigenspectra pair to be physically consistent with PC1 versus PC3 at late times (i.e., $t_{V_{\max}} = 10$, 15 days) because this pair of eigenspectra produces the highest SVM test score across all phases and the least amount of overlap of the 1σ contours of the PCA coefficients for the different SESN classes. We find that we can recreate the SNID labels of our data set. Furthermore, we find that the optimal phase ranges for classifying SESNe are $t_{V_{\max}} = 10 \pm 5$ days and $t_{V_{\max}} = 15 \pm 5$ days, as opposed to at maximum light ($t_{V_{\max}} = 0 \pm 5$ days). This is important in a future that, with the advent of LSST, will see an overwhelming number of SN discoveries, and a radical pressure on the urgency of spectroscopic follow-up for classification. Lowering the pressure on immediate follow-up for one type of transient (SESNe) alleviates pressure on the follow-up facilities altogether.

5.1. Classification in the PC1 versus PC3 Projection

Figure 5 shows the two-dimensional projection of our SESNe spectra onto the optimal eigenspectra pairs that maximally separate subclasses: PC1 versus PC3 for $t_{V_{\max}} = 5 \pm 5$, 10 ± 5, 15 ± 5 days, and PC1 versus PC5 for $t_{V_{\max}} = 0 \pm 5$ days (as we described in Section 4.2, PC5 at $t_{V_{\max}} = 0 \pm 5$ corresponds to PC3 in the later phase ranges). The colored regions illustrate the linear SVM decision boundaries. Boundaries for 50 different 70%–30% train-test splits of the data are shown, thus assessing the statistical robustness of the decision boundaries. The SVM test score, a measure of the accuracy of the classification, is indicated in each figure panel, including uncertainties generated from the 50 train-test splits. Colored ellipses in each panel represent the 1 standard deviation (1σ) contours of the PC coefficients for the different SESNe types. We have not included SNe Ib-pec (e.g., SN2007uy, SN2009er Modjaz et al. 2014) nor SNe Ic-pec (e.g., SN2005ek Drout et al. 2013) in the calculation of the ellipses (but we do show the data points of these peculiar subtypes).

Both SVM test score and the 1σ contours allow us to evaluate the success of our classifying scheme. The highest SVM test scores (.71 ± .10 and .70 ± .11) are achieved at $t_{V_{\max}} = 10 \pm 5$ and 15 ± 5 days, respectively. These scores, however, are statistically consistent at the 1σ or 2σ level with the lower test scores of the earlier phases. Nonetheless, the SESN classes are more compactly clustered and separated at later times as shown by the 1σ contours that are maximally

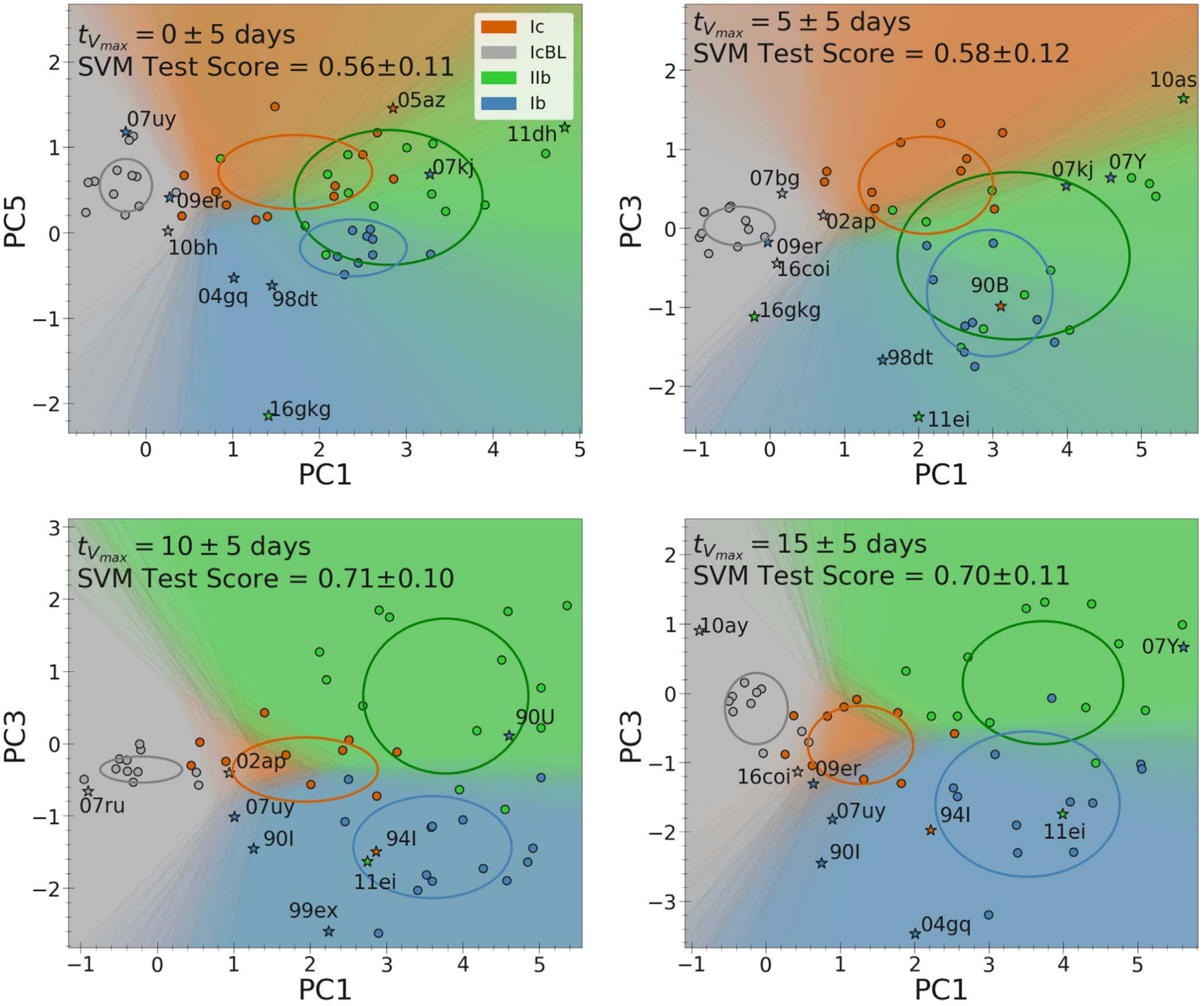

Figure 5. Each panel shows the SESNe classification regions and linear decision boundaries for each SVM train-test split of the data. Ellipses represent the 1 standard deviation contour of the PC coefficients for each SESN type (excluding the peculiar SNe SN2007uy, SN2009er, and SN2005ek). Outliers of more than 2 standard deviations from the mean are marked with stars. The phase range $t_{V_{\max}}$ is labeled in the upper left of each panel, along with the mean SVM test score. PC1 vs. PC3 provides the highest SVM test score for each phase range except $t_{V_{\max}} = 0 \pm 5$ where PC1 vs. PC2 has a slightly ($<1\sigma$) higher SVM test score but very similar 1σ contour and SVM region overlap. Upper left: ($t_{V_{\max}} = 0 \pm 5$ days). There is large overlap between the IIb (green), Ib (blue), and Ic (orange) 1σ contours, and between the SVM IIb and Ib region, and the IIb and Ic region (as indicated by the region boundaries changing significantly for different train-test splits and the colors bleeding into each other). Upper right: ($t_{V_{\max}} = 5 \pm 5$ days). As $t_{V_{\max}} = 0 \pm 5$ days, there is overlap between the IIb, Ib, and Ic 1σ contours, and the corresponding SVM regions. Lower left: ($t_{V_{\max}} = 10 \pm 5$ days). The IIb, Ib, Ic, and Ic-bl 1σ contours are completely separated. Each colored SVM region is well defined and stable for different train-test splits, and the SVM test score is highest. The Ic (orange) SVM region has collapsed and the IIb (green) SVM region has expanded. Lower tight: ($t_{V_{\max}} = 15 \pm 5$ days). SESNe type 1σ contours are well separated and the SVM regions are stable. The data used to create this figure are available.

separated at these later phases. Therefore, we find that the optimal time for classifying SESNe spectra is later than ($t_{V_{\max}} = 10 \pm 5$, 15 ± 5 days) rather than at or near peak ($t_{V_{\max}} = 0 \pm 5$, 5 ± 5 days).

PC1 is a poor choice of eigenspectrum for SESNe classification at early times because the He I5876 absorption feature in Ic, IIb, and Ib spectra has not had time to strengthen. In the phase ranges $t_{V_{\max}} = 10 \pm 5$ and $t_{V_{\max}} = 15 \pm 5$ days, PC1 and PC3 both become more effective at distinguishing between SESNe spectral types, with less overlap in the 1σ contours and a higher SVM test score. In particular, we find that the PC1 coefficients of SNe types IIb and Ib increase (while SNe Ic PC1 coefficients remain relatively unchanged) as phase increases. Since PC1 captures the strong feature at $\lambda \approx 5600$–5800 Å, which is due to He in SNe Ib and IIb, this behavior is consistent with Liu et al. (2016), which found that the pseudo-equivalent width (pEW) of the He I absorption features in SNe types IIb and Ib increases as a function of phase. Figure 5 also shows that as phase increases, PC3 becomes better at distinguishing between SNe types IIb (green region) and Ib (blue region). Specifically, the SNe type IIb PC3 coefficients systematically increase with increasing phase. Since PC3 captures the Hα and Hβ features, this behavior is consistent with the strengthening of the Hβ absorption feature in SNe IIb mean spectra shown in Liu et al. (2016).

The SNe Ic-bl region (gray) is reasonably well separated from the other SESNe types at all phases in Figure 5. However, note that the Ic-bl data and the corresponding 1 standard deviation contour is centered near the origin in every panel. Moreover, we find the PC coefficients of the SNe Ic-bl to be clustered around zero in every two-dimensional projection of the first five eigenspectra. This is expected because the SNe Ic-bl mean spectra do not have a strong absorption feature due to He I5876, even if it were highly broadened (Modjaz et al. 2016).

5.2. Transition Supernovae and Type Outliers in PCA Space

One major benefit of our work is that the PC coefficients of the SESNe in our sample are continuous, and therefore well suited for capturing the physical continuity of chemical abundances in SNe ejecta. This behavior is particularly useful for objectively identifying "transition" SNe, which often have debated classification in the literature due to spectra that resemble more than one SESN type. Our method also identifies outliers in a particular class that are extreme versions of the SESN type, but not "transition" SNe. In Figure 5 we label all SNe in each panel that are more than 2 standard deviation outliers and discuss them below.

5.2.1. Type Ib Outliers

Figure 5 shows two SNe Ib that are consistently strong outliers: SN2007uy and SN2009er. These two supernovae either appear within the SNe Ic-bl (gray) region or close to the SVM decision boundary separating the Ic-bl and Ib regions (note that if SN2007uy or SN2009er does not appear in a panel of Figure 5, it is because we have no spectra in the corresponding phase range). SN2007uy and SN2009er have been previously identified in the literature (Modjaz et al. 2014) as peculiar members of the Ib class. Modjaz et al. showed that SN2007uy and SN2009er have broader features at higher velocities than normal SNe Ib spectra, in agreement with our results. We also find that SN1990I, SN1998dt, and SN2004gq are consistent outliers toward the Ic-bl region, although to lesser degrees than SN2007uy and SN2009er. Elmhamdi et al. (2004) have previously identified SN1990I as having high velocity features atypical of a normal SN Ib, and Modjaz et al. (2014) show that SN2004gq and SN1990I both have high absorption velocity He features compared to other SNe Ib spectra. The outliers SN1990I, SN1998dt, and SN2004gq may form a continuum of SN Ib spectra with higher than normal Doppler shifts, while SN2007uy and SN2009er indicate the possibility for a continuum of SNe Ib spectra with varying amounts of line blending. SN1999ex was initially classified as an SN Ic, then changed to an SN Ib/c due to moderate He I absorption features (Hamuy et al. 2002). More recently, SN1999ex has been classified as an SN Ib (Modjaz et al. 2014). We identify SN1999ex as an outlier in multiple 2D projections of PCA space, indicating that it is not a standard SN Ib nor a standard SN Ic.

We also identify SN1990U, SN2007kj, and SN2007Y as outliers in Figure 5. SN1990U (found in the green SN IIb region) has previously been considered as an SN Ic (Matheson et al. 2001) and more recently as an SN Ib (Modjaz et al. 2014). Although we identify SN1990U as an outlier SN Ib in the PC1 versus PC3 projection, in the other projections it is an SN Ib, and in no projection is SN1990U located in the standard SN Ic region. Therefore, our results support the reclassification of SN1990U as an SN Ib by Modjaz et al. (2014). SN2007kj was previously classified as an SN Ib/c "transition" object (Leloudas et al. 2011) and more recently as an SN Ib (Modjaz et al. 2014). We find that SN2007kj would be considered a strong outlier as an SN Ic in every 2D projection of the first five eigenspectra, while it is consistent with being a standard SN Ib in multiple 2D projections (not shown) other than PC1 versus PC3. Therefore we support the reclassification of SN2007kj as an SN Ib by Modjaz et al. (2014). SN2007Y has been classified both as an SN IIb (Folatelli et al. 2014) and an SN Ib (Liu et al. 2016). In the PC1 versus PC3 2D projection, we find that SN2007Y falls in the IIb region, consistent with Folatelli et al. (2014), who argued that SN2007Y is an SN IIb due to the strength and velocity of the He I5876 feature. However, in another 2D projection (not shown), namely, PC1 versus PC4 (strong Hα and Hβ features) at phases $t_{V_{\max}} = 5 \pm 5$, 15 ± 5 days, we find that SN2007Y falls in the Ib region, in agreement with Liu et al. (2016), who found that the H feature evolution of SN2007Y was consistent with SN Ib spectra. Thus, our classification method captures the debate over the correct type for SN2007Y.

5.2.2. Type IIb Outliers

In Figure 5 we label the following outlier SNe IIb: SN2010as, SN2011ei, and SN2016gkg. At early times, SN2010as appears on the decision boundary between types Ic (orange) and IIb (green), which is consistent with Folatelli et al. (2014), who found that SN2010as exhibits weaker than normal He features at early times, in addition to weak H features. SN2011ei is a strong outlier in the PC1 versus PC3 2D projection. Milisavljevic et al. (2013) showed that SN2011ei evolves quickly, losing its H features within a week after the V-band maximum, to resemble a type Ib spectrum characterized by helium features. Figure 5 illustrates this evolution, with SN2011ei initially a standard IIb at phase $t_{V_{\max}} = 0 \pm 5$ days, then subsequently moving to the Ib region. However, Liu et al. (2016) showed that the Hα equivalent width evolves differently for type IIb and Ib spectra (including SN2011ei), so SN2011ei is distinguishable as an SN IIb even at late times. When we consider the PC1 versus PC4 (strong Hα feature) 2D projections (not shown) at the later phase ranges $t_{V_{\max}} = 5 \pm 5$, 10 ± 5, 15 ± 5 days, we find SN2011ei to be consistently within the IIb (green) region in agreement with Liu et al. (2016). SN2016gkg is classified as an SN IIb due to its Hα absorption, but Tartaglia et al. (2017) showed that SN2016gkg exhibits stronger than normal helium features even at early times, similar to an SN Ib. Figure 5 captures this behavior, showing SN2016gkg as a strong outlier in the Ib (blue) region at $t_{V_{\max}} = 0 \pm 5$ days, but a more normal SN Ib in other 2D projections (not shown).

5.2.3. Type Ic/Ic-bl Outliers

We identify three Ic outliers, SN1990B, SN1994I, SN2005az, and six Ic-bl outliers, SN2002ap, SN2007bg, SN2007ru, SN2010ay, SN2010bh, and SN2016coi in Figure 5. SN1990B is currently considered an SN Ic; however, it was initially classified as an SN Ib (Clocchiatti et al. 2001), which is consistent with our results in Figure 5. SN1994I is one of only a few SN Ic with many spectra taken over a range of wavelength regimes (e.g., Filippenko et al. 1995; Richmond et al. 1996; Immler et al. 1998), and it is considered a prototypical SN Ic. However, we find that SN1994I is considered an outlier in many 2D PCA projections, at multiple phases, as illustrated in Figure 5. Our results indicate that SN1994I may not be a prototypical SN Ic, confirming the spectroscopic analysis of

Modjaz et al. (2016) and the photometric analysis of Drout et al. (2011) and Bianco et al. (2014). SN2005az was initially classified as both an SN Ic (Aldering et al. 2005) and an SN Ib (Quimby et al. 2005). Recently SN2005az has been classified as an SN Ic (Kelly & Kirshner 2012) using SNID based on the updated SESNe library from Modjaz et al. (2014) and Liu et al. (2016). We find that SN2005az is inconsistent with being an SN Ib in the majority of 2D projections, and when it is consistent with belonging to the Ib or IIb class, this is due to large overlap of the Ic and Ib/IIb regions at $t_{V_{\max}} = 0 \pm 5$ days. Meanwhile, there are some PCA 2D projections (PC2 versus PC4; not shown) where SN2005az is located within the SN Ic one standard deviation contour, so we support the classification of SN2005az as an SN Ic.

SN2002ap is claimed to be a relatively low energy SN Ic-bl compared to normal SNe Ic-bl events (Mazzali et al. 2002) and has been classified as a normal SN Ic from radio observations (Berger et al. 2002). We find that SN2002ap is indeed a potential transition object between the Ic and Ic-bl regions in Figure 5. Although SN2007bg is identified as an outlier at phase $t_{V_{\max}} = 5 \pm 5$ days, it is well within the Ic-bl SVM region (gray), and it no longer fulfills the outlier criterion at later phases, in agreement with the literature view that SN2007bg is a standard SN Ic-bl (Young et al. 2010). Similarly, although SN2007ru is marked as an outlier in the lower left panel of Figure 5, it is well within the SN Ic-bl SVM region and considered a standard SN Ic-bl (Sahu et al. 2009). SN2010bh is considered a standard SN Ic-bl, although with slightly higher inferred explosion energy than other standard Ic-bl SNe (Chornock et al. 2010). At late times (bottom right) of Figure 5, we find that SN2010ay is a strong SN Ic-bl outlier well within the Ic-bl region. SN2010ay is a particularly interesting Ic-bl because it has been proposed that SN2010ay was associated with an off-axis low-luminosity gamma-ray burst, due to its high absorption velocity, high peak luminosity, and low metallicity (Sanders et al. 2012), combined with a lack of observed gamma-rays. SN2016coi has broad spectral features in addition to a strong absorption feature generally attributed to He I in the literature (Prentice et al. 2018), setting it apart from normal Ic-bl SNe. We find that SN2016coi is located right at the SVM boundary between the Ic-bl and Ib regions consistent with SN2016coi being similar to the SN Ib class.

6. Summary and Future Work

In this work, we have shown that PCA is a useful tool as a first step toward a data-driven classification method for SESNe types. We used multiclass linear SVMs to explore different projections of SESNe spectra onto eigenspectra and found that the SESNe types are more distinguishable in the later phase ranges $t_{V_{\max}} \approx 10 - 15$ days relative to the V-band maximum, rather than at peak light. We recommend that spectral follow-up of ZTF and LSST supernovae take these considerations into account. In addition, our classification method naturally provides a continuous, quantifiable method for characterizing "transition" SNe based on distance to class boundaries or centroids. We showed that our classification method identified both "transition" SNe and SNe with debated types previously identified in the literature, and we interpreted these SNe using our PCA eigenspectra and our SVM classification regions.

PCA is clearly a promising dimensionality reduction tool for SESNe, and there are many future projects that would use the work presented here as a starting point. In particular, the probability of a supernova's membership in one of the SESNe types could be calculated using the distance from its PCA projection to an SVM decision boundary. This provides a quantitative understanding of "transition" SNe like the type Ibc's, and should especially be explored as a function of phase.

We would like to thank Wolfgang Kerzendorf, Tyler Pritchard, David Hogg, Coryn Bailer-Jones, Markus Kromer, and Fritz Roepke for useful discussions. M.M. and the SNYU group are supported by the NSF CAREER award AST-1352405, by the NSF award AST-1413260, and by a Humboldt Faculty Fellowship.

This work is based (in part) on observations collected at the European Organisation for Astronomical Research in the Southern Hemisphere, Chile as part of PESSTO (the Public ESO Spectroscopic Survey for Transient Objects Survey), ESO program 188.D-3003, 191.D-0935.

ORCID iDs

Marc Williamson https://orcid.org/0000-0003-2544-4516
Maryam Modjaz https://orcid.org/0000-0001-7132-0333
Federica B. Bianco https://orcid.org/0000-0003-1953-8727

References

Aldering, G., Lee, B. C., Loken, S., et al. 2005, ATel, 451, 1
Bellm, E. C., Kulkarni, S. R., Graham, M. J., et al. 2019, PASP, 131, 018002
Berger, E., Kulkarni, S., & Chevalier, R. 2002, ApJL, 577, L5
Bianco, F., Modjaz, M., Hicken, M., et al. 2014, ApJS, 213, 19
Bianco, F. B., Koonin, S. E., Mydlarz, C., & Sharma, M. S. 2016, in Proc. III ACM Int. Conf. Systems for Energy-Efficient Built Environments, BuildSys 2016 (New York: ACM), 61
Blondin, S., & Tonry, J. L. 2007, ApJ, 666, 1024
Bufano, F., Pignata, G., Bersten, M., et al. 2014, MNRAS, 439, 1807
Cano, Z., Wang, S.-Q., Dai, Z.-G., & Wu, X.-F. 2017, AdAst, 2017, 8929054
Childress, M. J., Tucker, B. E., Yuan, F., et al. 2016, PASA, 33, e055
Chornock, R., Berger, E., Levesque, E. M., et al. 2010, arXiv:1004.2262
Clocchiatti, A., Suntzeff, N. B., Phillips, M. M., et al. 2001, ApJ, 553, 886
Clocchiatti, A., Wheeler, J., Phillips, M., et al. 1997, ApJ, 483, 675
Cormier, D., & Davis, T. M. 2011, MNRAS, 410, 2137
Dessart, L., & Hillier, D. J. 2010, MNRAS, 405, 2141
Drout, M., Milisavljevic, D., Parrent, J., et al. 2016, ApJ, 821, 57
Drout, M., Soderberg, A. M., Mazzali, P., et al. 2013, ApJ, 774, 58
Drout, M. R., Soderberg, A. M., Gal-Yam, A., et al. 2011, ApJ, 741, 97
Elmhamdi, A., Danziger, I. J., Cappellaro, E., et al. 2004, A&A, 426, 963
Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. 2008, Journal of Machine Learning Research, 9, 1871
Filippenko, A. V. 1997, ARA&A, 35, 309
Filippenko, A. V., Barth, A. J., Matheson, T., et al. 1995, ApJL, 450, L11
Filippenko, A. V., Matheson, T., & Ho, L. C. 1993, ApJL, 415, L103
Folatelli, G., Bersten, M. C., Kuncarayakti, H., et al. 2014, ApJ, 792, 7
Fremling, C., Sollerman, J., Kasliwal, M., et al. 2018, A&A, 618, A37
Fremling, C., Sollerman, J., Taddia, F., et al. 2016, A&A, 593, A68
Gal-Yam, A. 2017, in Handbook of Supernovae, ed. A. W. Alsabti & P. Murdin (Berlin: Springer), 1
Hamuy, M., Maza, J., Pinto, P. A., et al. 2002, AJ, 124, 417
Howell, D. A., Sullivan, M., Perrett, K., et al. 2005, ApJ, 634, 1190
Immler, S., Pietsch, W., & Aschenbach, B. 1998, A&A, 336, L1
Ivezic, Z., Tyson, J., Abel, B., et al. 2008, arXiv:0805.2366
Jolliffe, I. 2011, International Encyclopedia of Statistical Science (Berlin: Springer), 1094
Kelly, P. L., & Kirshner, R. P. 2012, ApJ, 759, 107
Leloudas, G., Gallazzi, A., Sollerman, J., et al. 2011, A&A, 530, A95
Liu, Y., & Modjaz, M. 2014, arXiv:1405.1437
Liu, Y.-Q., Modjaz, M., Bianco, F. B., & Graur, O. 2016, ApJ, 827, 90

Matheson, T., Filippenko, A. V., Li, W., Leonard, D. C., & Shields, J. C. 2001, AJ, 121, 1648

Mazzali, P., Deng, J., Maeda, K., et al. 2002, ApJL, 572, L61

Milisavljevic, D., Margutti, R., Soderberg, A. M., et al. 2013, ApJ, 767, 71

Minkowski, R. 1941, PASP, 53, 224

Modjaz, M. 2011, RvMA, 23, 11

Modjaz, M., Blondin, S., Kirshner, R. P., et al. 2014, AJ, 147, 99

Modjaz, M., Liu, Y. Q., Bianco, F. B., & Graur, O. 2016, ApJ, 832, 108

Nicholl, M., Berger, E., Blanchard, P. K., Gomez, S., & Chornock, R. 2019, ApJ, 871, 102

Pearson, K. 1901, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2, 559

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825

Prentice, S., Ashall, C., Mazzali, P., et al. 2018, MNRAS, 478, 4162

Prentice, S. J., & Mazzali, P. A. 2017, MNRAS, 469, 2672

Quimby, R., Mondol, P., Hoeflich, P., Wheeler, J. C., & Gerardy, C. 2005, IAUC, 8503, 1

Quimby, R. M., De Cia, A., Gal-Yam, A., et al. 2018, ApJ, 855, 2

Richmond, M. W., Van Dyk, S., Ho, W., et al. 1996, AJ, 111, 327

Sahu, D., Anupama, G., Chakradhari, N., et al. 2018, MNRAS, 475, 2591

Sahu, D., Tanaka, M., Anupama, G., Gurugubelli, U. K., & Nomoto, K. 2009, ApJ, 697, 676

Sanders, N. E., Soderberg, A. M., Valenti, S., et al. 2012, ApJ, 756, 184

Sasdelli, M., Hillebrandt, W., Aldering, G., et al. 2014, MNRAS, 447, 1247

Shivvers, I., Filippenko, A. V., Silverman, J. M., et al. 2018, MNRAS, 482, 1545

Smartt, S. J., Valenti, S., Fraser, M., et al. 2015, A&A, 579, A40

Sun, F., & Gal-Yam, A. 2017, arXiv:1707.02543

Szalai, T., Vinkó, J., Nagy, A. P., et al. 2016, MNRAS, 460, 1500

Taddia, F., Fremling, C., Sollerman, J., et al. 2016, A&A, 592, A89

Taddia, F., Sollerman, J., Fremling, C., et al. 2018, arXiv:1811.09544

Takaki, K., Kawabata, K. S., Yamanaka, M., et al. 2013, ApJL, 772, L17

Tartaglia, L., Fraser, M., Sand, D., et al. 2017, ApJL, 836, L12

Van Dyk, S. D., Zheng, W., Brink, T. G., et al. 2018, ApJ, 860, 90

Vapnik, V. 1998, Nonlinear Modeling (Berlin: Springer), 55

Woosley, S., & Bloom, J. 2006, ARA&A, 44, 507

Woosley, S. E., Heger, A., & Weaver, T. A. 2002, RvMP, 74, 1015

Yaron, O., & Gal-Yam, A. 2012, PASP, 124, 668

Young, D., Smartt, S., Valenti, S., et al. 2010, A&A, 512, A70