

**A Metric Space
for Series of Spectra
of Type Ia Supernovae**

DISSERTATION

Michele Sasdelli

TECHNISCHE UNIVERSITÄT MÜNCHEN

MAX-PLANCK-INSTITUT FÜR ASTROPHYSIK

**A Metric Space
for Series of Spectra
of Type Ia Supernovae**

Michele Sasdelli

Vollständiger Abdruck der von der Fakultät für Physik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende: Univ.-Prof. Dr. Nora Brambilla

Prüfer der Dissertation:

1. Hon.-Prof. Dr. Wolfgang Hillebrandt
2. Univ.-Prof. Shawn Bishop, Ph.D.

Die Dissertation wurde am 20.04.2015 bei der Technischen Universität München eingereicht und durch die Fakultät für Physik am 11.06.2015 angenommen.

Abstract

A metric space for the study of large samples of spectral series of type Ia supernovae is introduced. The method is built to be intrinsically independent from dust extinction and estimates of the distance and its validity is confirmed by checking a number of known relations. The resulting metric space is ideal for a number of applications. Some are investigated in detail: assess spectral series of models, calibrate near infrared luminosity, and study the extinction-reddening relation.

Zusammenfassung

In der vorliegenden Dissertation wird ein metrischer Raum zur Untersuchung von großen Sätzen spektraler Serien von Typ-Ia-Supernovae eingeführt. Diese Methode ist so konstruiert, dass sie unabhängig von Extinktion durch Staub und einer Entfernungsabschätzung ist. Die Gültigkeit der Methode wurde anhand mehrerer bereits bekannter Relationen überprüft. Der gefundene metrische Raum eignet sich für eine Vielzahl von Anwendungen, von denen einige genauer untersucht wurden: die Bewertung von Modell-Spektralserien, die Kalibration von Nah-Infrarot-Spektroskopie, und die Untersuchung der Extinktions-Rotverschiebungs-Relation.

Contents

1	Introduction	1
1.1	The Observed Properties of Supernovae	1
1.2	The Explosion Mechanisms and Progenitors of Supernovae	6
1.3	Observational Efforts	8
1.4	Statistical Methods for Type Ia Supernovae	9
2	The Method	11
2.1	Filtering the Spectra	11
2.1.1	Weighted Savitzky-Golay Filter	11
2.1.2	Improving the Filtering	12
2.2	Expectation Maximization Principal Component Analysis	12
2.3	Error Budget	14
2.4	Optimizing Information Extraction	15
2.4.1	Derivative Spectroscopy	15
2.4.2	Complete Spectral Sequences	16
2.5	The Partial Least Square Analysis	17
3	Applications	19
3.1	The Nearby Supernova Factory	19
3.2	Publicly Available Spectra	24
4	Principal Components Interpretation and Metric Space Comparison	27
4.1	Principal Components	27
4.2	Metric Spaces	29
4.3	High Velocity Features	34
4.4	Spectral Series and EMPCA on the Publicly Available Spectra	37
5	Comparison with Discrete Observables	43
5.1	Measurement of Discrete Observables	44
5.2	Results from Partial Least Square	45
5.3	Spectroscopic Observables in Derivative Principal Component Space	47
5.4	Photometric Observables in Derivative Principal Component Space	49

5.5	Infrared Light Curve Properties and Spectral Properties	52
5.5.1	Infrared Second Maximum	54
5.5.2	Type Ia Supernovae are “Standardizable” Candles in the Near Infrared	54
6	Comparison with Models	59
6.1	Explosion Scenarios	60
6.2	Models in the Principal Component Space	62
6.2.1	$\Delta m_{15}(B)$	65
6.2.2	$B - V$ Color	67
6.2.3	B Magnitudes	70
7	Multivariate Partial Least Square as a Light Curve Predictor	73
7.1	Interpolate the Photometry with Gaussian Processes	74
7.2	Multivariate Partial Least Square	75
7.2.1	The Algorithm of Multivariate Partial Least Square	75
7.3	Predicting Light Curves and Color Curves from the Spectra	77
7.3.1	Predicting the Light Curve from the Spectra	79
7.3.2	Predicting the Color Curve from the Spectra	81
7.4	Studying the Extinction-Reddening Relation	85
8	Conclusions and Outlook	91
A	Appendix	95
A.1	Cross-Validation	95
A.2	Line Velocities and Pseudo Equivalent Width Calculations	98
A.3	The reconstructions in the derivative space	98

List of Figures

1.1	Spectral variability of different SNe types	2
1.2	Light curve variability of different SNe types	3
1.3	Light curve variability of within SN Ia	4
1.4	Spectral variability of SN Ia	5
1.5	Spectral evolution of a normal SN Ia	7
3.1	Calculating the derivative	21
3.2	The input matrix	22
4.1	The eigenvectors of flux PCA	28
4.2	The eigenvectors of derivative PCA	28
4.3	Flux PCA scatter plot	30
4.4	Flux PCA scatter plot – PC2 and PC3	31
4.5	derivative PCA scatter plot	32
4.6	Eigenvalues of the covariance of the flux	34
4.7	Eigenvalues of the covariance of the derivative	34
4.8	Reconstruction of the spectra	35
4.9	HVF in the PCA space	38
4.10	Dimensionality of the HVF subspace	39
4.11	Scatter plot of the public sample	40
4.12	Scatter plot of the public sample – PC1 and PC3	41
4.13	Reconstruction of SN2008Z	42
5.1	PLS directions	46
5.2	Si II 6355 Å velocity	48
5.3	S II 5640 Å velocity	48
5.4	Si II 5972 Å EW	48
5.5	Si II 6355 Å EW	48
5.6	<i>B</i> -band magnitude	50
5.7	<i>V</i> -band magnitude	50
5.8	<i>B</i> – <i>V</i> -color	50

5.9	SALT2 c	50
5.10	$\Delta m_{15}(B)$	51
5.11	SALT2 X1	51
5.12	The NIR light curve	53
5.13	The epoch of the J -band maximum	55
5.14	Calibrate the luminosity in the H -band	57
5.15	H -band luminosity and the spectra.	58
5.16	The directions of $\Delta m_{15}(B)$ and H -mag	58
6.1	Models in the PCA space	63
6.2	Models in the PCA space – PC1 and PC3	64
6.3	Models in the $\Delta m_{15}(B)$ -PLS space	66
6.4	Models in the $B - V$ -PLS space	68
6.5	Models in the B -band PLS space	71
7.1	Gaussian Processes	76
7.2	Multivariate PLS	78
7.3	B -band light curve predicted by the spectra	80
7.4	Dust extinction in the B -band	82
7.5	The Phillips relationship	83
7.6	The epoch of the $B - V$ color maximum	84
7.7	$B - V$ color curve predicted by the spectra	85
7.8	Dust extinction of the $B - V$ color	86
7.9	Reddening law	88
7.10	Reddening law (zoom)	89
A.1	K-folding – scatter plot	96
A.2	K-folding – x_1	97
A.3	Calculation of EW	99
A.4	Reconstructions of the derivatives	100

List of Tables

5.1	PLS latent variables	45
5.2	Quality of the predictions	47
A.1	K-folding – quality of the predictions	98

List of Constants

Physical Constants:¹

gravitational constant	G	$6.673\,84 \times 10^{-8} \text{ cm}^3 \text{ g}^{-1} \text{ s}^{-2}$ $6.673\,84 \times 10^{-11} \text{ m}^3 \text{ kg}^{-1} \text{ s}^{-2}$
speed of light	c	$2.997\,924\,580 \times 10^{10} \text{ cm s}^{-1}$ $2.997\,924\,580 \times 10^8 \text{ m s}^{-1}$

Astrophysical Quantities:²

solar mass	M_{\odot}	$1.989 \times 10^{33} \text{ g}$ $1.989 \times 10^{30} \text{ kg}$
solar radius	R_{\odot}	$6.955\,08 \times 10^{10} \text{ cm}$ $6.955\,08 \times 10^8 \text{ m}$
solar luminosity	L_{\odot}	$3.845 \times 10^{33} \text{ erg s}^{-1}$ $3.845 \times 10^{26} \text{ W}$
solar absolute B magnitude	M_B^{\odot}	5.47

¹CODATA 2010 recommended values (Mohr et al., 2012).

²Values taken from Allen's Astrophysical Quantities, 4th Edition (Cox, 2000).

This thesis is dedicated to the memory of my father and his art.
Thank you for teaching me to always aim at high things.

Acknowledgments

First of all I want to thank my supervisor Wolfgang Hillebrandt for all his support during this period of my life. Your capability of seeing ahead the right route in uncharted waters is inspiring. A big thank to my co-supervisor Paolo Mazzali. You framed my view of radiation transport and that showed me the right direction at the beginning of this journey. A special thank you is for Emille Ishida, the discussions with you have been inspiring and invaluable to clarify the ideas in my mind. You always find in no time the right tool for the job.

I want to thank Alex Kim and Greg Aldering and all the SNfactory collaboration. Thanks to you we improved on every rough edge of the method. And I want to thank to all the people providing data and ideas that have been the basis of this work, this include Markus Kromer, Steven Bailey, Stuart Sim, Ivo Seitzzahl, Michael Fink, Michael Klauser, Rüdiger Pakmor, and Friedrich Roepke. Thank you Philipp Edelman for all the technical support and for being a good friend. I thank all the python, numpy, scipy and scikit-learn communities for the high-quality free software they made available.

A big thank you to my parents, you always supported me by being incredibly proud whether I was doing good or bad alike. Thanks to all my friends. I own you a lot. Especially thank you Dominic, you have been there when I mostly needed you. And thank you to my girlfriend Sara, I think about this thesis as our achievement.

Chapter 1

Introduction

The aim of this thesis is to increase the understanding of Type Ia Supernovae (SNe Ia). This class of supernovae are among the most luminous transients in our local Universe. They hold a good degree of spectroscopic and photometric homogeneity. Their homogeneity allowed to use them as standard candles to measure the distance of remote galaxies and to study the expansion flow of the Universe. This led to the Nobel Prize awarded discovery of the Dark Energy, a mysterious repulsive fifth force that acts on the largest scales.

In the next sections we introduce the spectroscopic and luminosity behavior of SN Ia in the more general context of supernovae. We discuss the possible explosion scenarios and progenitor systems. Then, we present the current observational efforts to unravel their nature. Finally, we motivate and discuss the topic of the thesis, a statistically based approach to study their spectra and to extract from them information concerning the physics of the explosion.

1.1 The Observed Properties of Supernovae

SN Ia are not the only class of supernovae. The different subclasses are distinguished by their spectroscopic and photometric behaviour. Fig. 1.1 shows the classic classification scheme of SNe, based on the elements that show up in their optical spectra near maximum light. Type II SNe show hydrogen in their optical spectra and Type I do not. Type I SNe are named Ib if they show He lines. Finally, a SN is promoted to a Ia by the presence of strong lines of intermediate mass elements (IME), such as Si, S, and Ca.

Fig. 1.2 shows the light curves of many different SN types. From this figure it is evident that core-collapse SNe have a large diversity of maximum luminosity, rise time, and behaviour during the declining phase. In contrast, light curves of SNe Ia are remarkably uniform. Fig. 1.3 shows the light curves of a representative sample of SNe Ia. A careful observer will notice that the fainter objects are also characterized by a quicker decline of

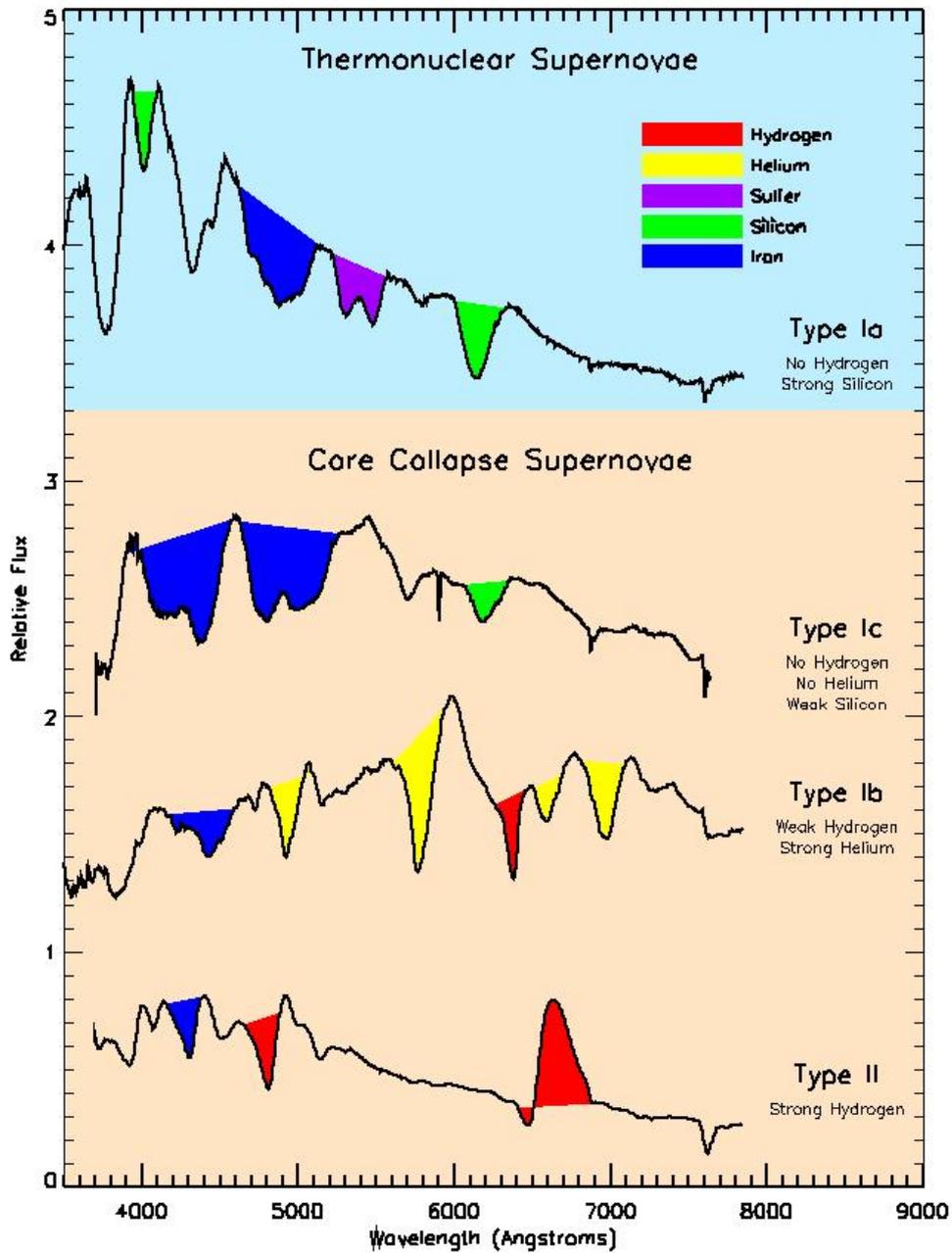


Figure 1.1: The classification scheme of supernovae is based on maximum-light visible spectra. Core-collapse SNe, originating from massive stars, offer a large diversity of spectroscopic characteristics (figure from <http://supernova.lbl.gov/~dnkasen/tutorial/>).

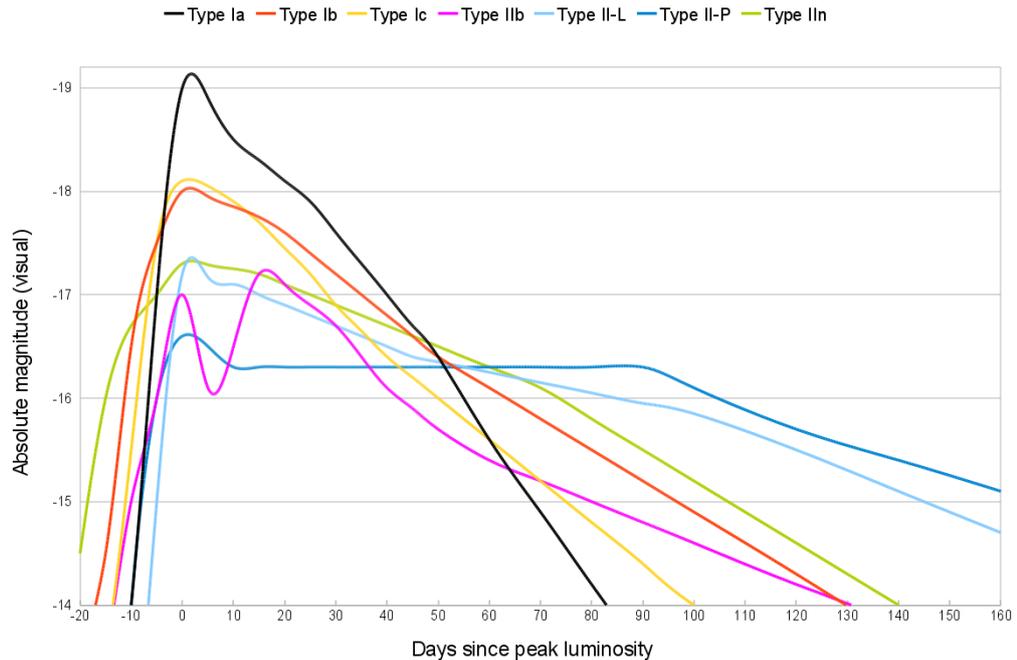


Figure 1.2: The light curves of supernovae show a large diversity of characteristics (magnitudes correspond to luminosity on a logarithmic scale). Type Ia are thought to be thermonuclear explosion of white-dwarfs. All other types: collapse of a massive star to a neutron star or a black hole (see also Sec. 1.2). The figure is taken from http://commons.wikimedia.org/wiki/File:Comparative_supernova_type_light_curves.png.

their luminosity. This relation, first discovered by Barbon et al. (1973); Pskovskii (1977, 1984); Phillips (1993), is the basic ingredient that allows to “standardize” the luminosity of these objects and to use them as indicators for relative distances. After the discovery of a relation between their light curve shape and luminosity and of the luminosity-color relation (e.g. Riess et al., 1996a; Tripp, 1998) they have served as “standardizable candles” and distance indicators in cosmology.

SNe Ia spectra, albeit quite homogeneous, exhibit a non-negligible diversity of spectral features (e.g. Benetti et al., 2005; Branch et al., 2006; Hachinger et al., 2006; Wang & Han, 2012). This diversity is clearly seen in Fig. 1.4, where we compare the typical subclasses of SN Ia at maximum light. In this figure we also present the lines most commonly used for classification and analysis.

From the selection of spectra shown in Fig. 1.4, it is already clear that within the SN Ia class there is a large diversity in spectral properties. From top to bottom the SNe are roughly ordered by luminosity. At the top are shown SN2000cx and SN1999aa, two SNe belonging to the 1991T-like subclass, characterized by “hot” spectra, with an high

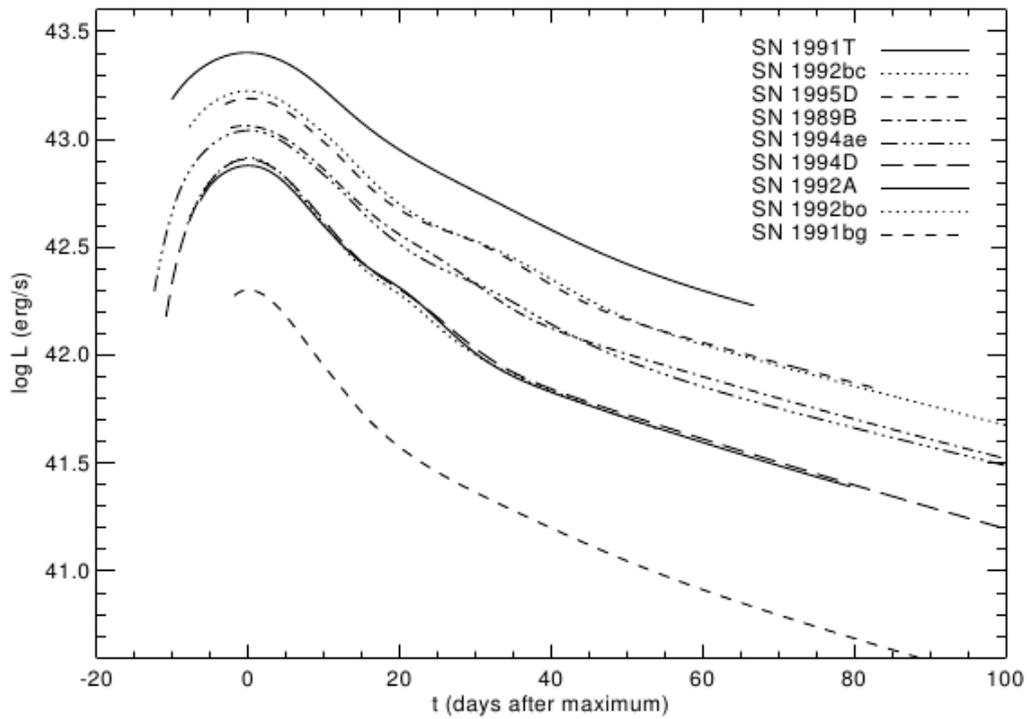


Figure 1.3: This is a representative sample of SNe Ia bolometric light curves. They are very well behaved. The maximum luminosity clearly anti-correlates with the decline-rate (figure from Contardo et al. (2000)).

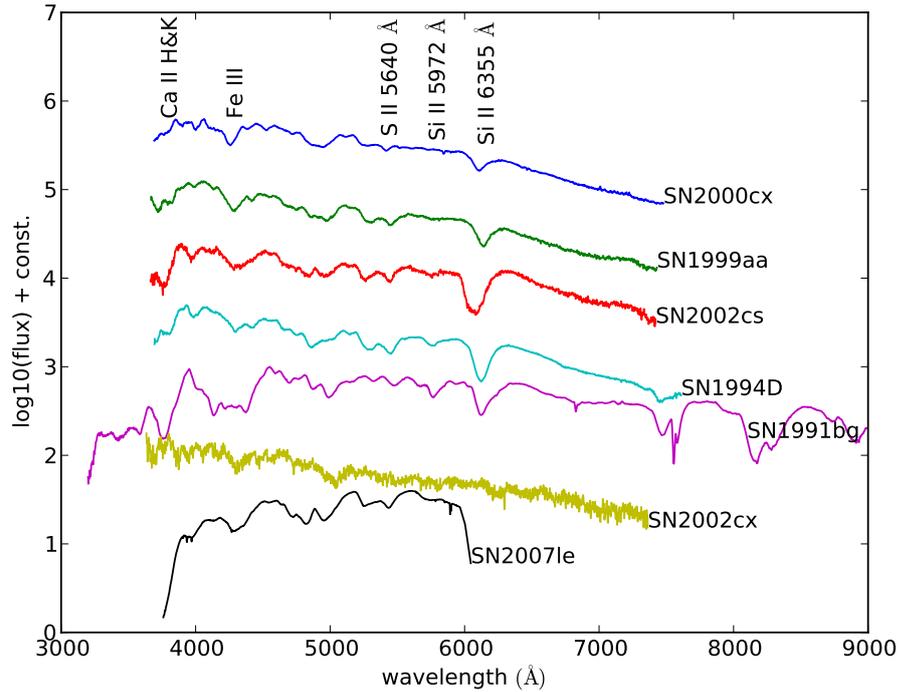


Figure 1.4: Spectra at maximum light of a selection of SN Ia. From the top, the first two object are 1991T-like SNe; then we have a supernova with a high-velocity of Si (more blueshifted features); SN1994D, the “prototypical” normal SN Ia; SN1991bg, that gave the name to a class of faint SNe; SN2002cx, another class of faint SN Ia are named after it; finally, the reddened SN 2007le, that without reddening would be very similar to SN2002cs.

ionization, and the presence of strong Fe III lines. The Si II 6355 Å line is weaker than in normal SN Ia and the Si II 5972 Å is absent. Also, the Ca II H&K line is usually weak or absent. Supernovae with these spectral features are commonly called “SN1991T-like” after the first event of this kind ever discovered. Then we have SN2002cs, a high-velocity Si supernova, characterized by broad and blueshifted features of IME such as Si II and S II lines. The next is SN1994D, the “prototypical” normal SN Ia where all the features considered are visible at a “moderate” velocity. The last three are SNe significantly fainter than the average, but for different reasons. SN1991bg and SN2002cx are the first discovered representatives of two different classes of faint SN Ia. 1991bg-like SNe have “cold” spectra, with normal velocities, and 2002cx-like SNe have very “hot” spectra, but with very low velocities. SN2007le, on the other hand, is a SN intrinsically very similar to SN2002cs, but with a significant amount of dust along the line of sight. This dust is responsible for decreasing its luminosity and, absorbing more on the blue edge of the spectrum, changing its color. The problem is even more complicated, because there are no sharp boundaries between these classes, and there are many SNe with intermediate properties. This sample selection is explicative of some of the typical problems of supernova spectroscopy, such as irregular wavelength coverage and diverse signal-to-noise ratios. In addition, the spectrum of a SN Ia varies heavily with time, and this variation contains valuable information about the physics of the object. Studying their spectral differences is a promising way to shed some light on questions regarding their nature.

Fig. 1.5 shows how the spectrum of a normal SN Ia changes with time. As the volume of the ejecta expands, the density decreases and the gas become more and more transparent. Progressively, the inner layers of the SN become “visible” and affect the characteristics of the spectrum. This is why it is important to consider the information of the spectral variability in spectroscopic analyses. The red area in Fig 1.5 is what is included in the statistical analysis of the next chapters. We tried to make it as large as possible in the wavelength and time domain, but at the same time we are limited by the incompleteness of the data. Not all SNe are observed as well as SN2011fe, one of the closest SNe in a quarter of a century.

1.2 The Explosion Mechanisms and Progenitors of Supernovae

Type II, Type Ib and Ic SNe are known to be the final bursts of massive stars ($> 8M_{\odot}$) undergoing gravitational collapse of the core and, therefore, called core-collapse SNe. They obtain their energy from the increase of gravitational binding of their final state: a neutron star or a black hole. The large diversity in progenitor mass, progenitor type, and type of the explosion is reflected in the large diversity of their spectra and light curves.

SN Ia, on the other hand, appear to be a rather homogeneous group, both photometrically and spectroscopically. Albeit being relatively common, quite homogeneous and

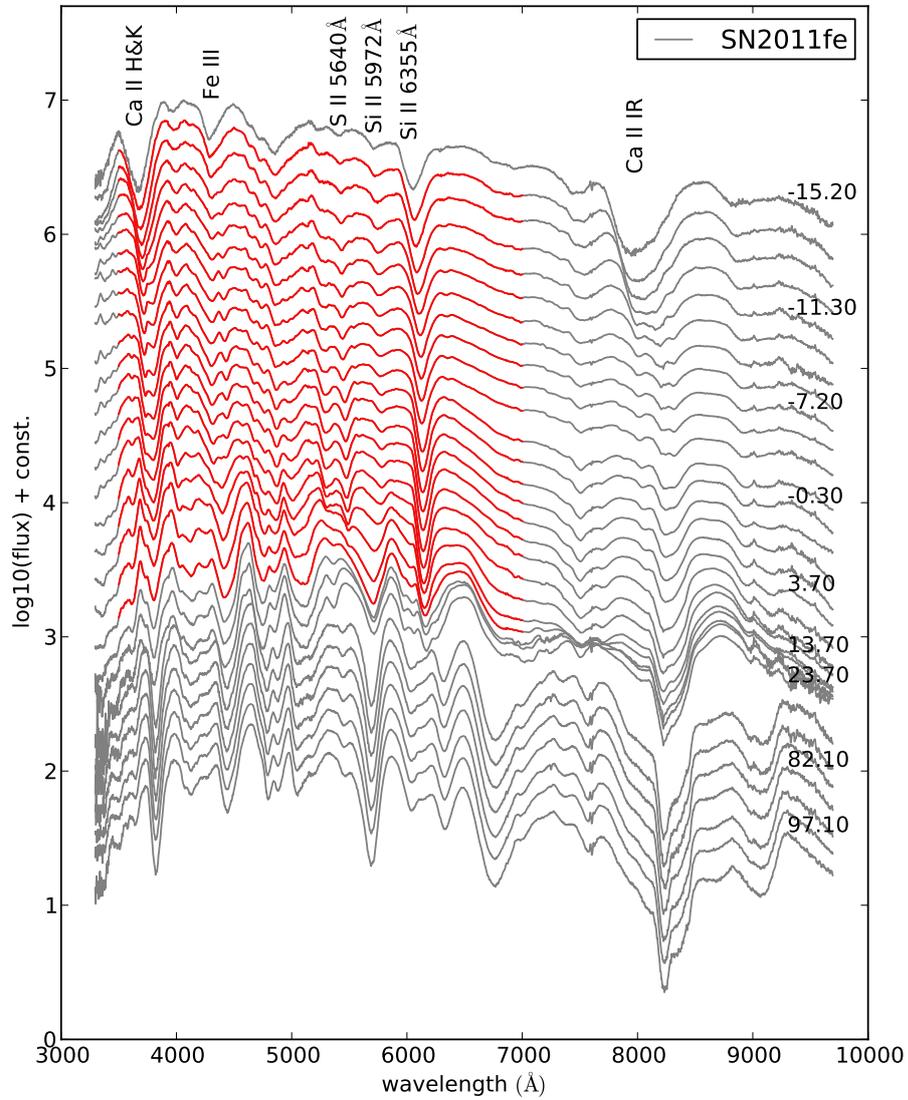


Figure 1.5: The figure shows the spectral evolution of the normal SN Ia, SN2011fe. The evolution of the spectrum have been followed from two weeks before maximum up to a hundred days after. The “square” delineated in red highlights the spectral variability that we include in our statistical analysis (chapter 3).

significantly bright (compared to other SN types), their precise nature is not clearly determined. Differently from other SN types, neither the progenitor nor a compact remnant have ever been identified with certainty. This absence suggests that the progenitors are rather faint and do not have the necessary mass to produce a compact remnant such as a neutron star.

The more uniform behaviour of the light curves of SN Ia is easily explained if we assume that the light comes from radioactive decay. The light curves first decline rapidly for a few weeks, followed by a slower decline over several months to a year. These timescales are driven by the radioactive heating of the beta decay chain of ^{56}Ni and ^{56}Co :



From a tenth up to one solar mass (M_{\odot}) of ^{56}Ni are synthesized during the explosion of SNe Ia, and this drives the luminosity of the object. The fast rise time of these objects suggest that the total mass of the ejecta is relatively uniform and quite small. White dwarfs are stars that exhausted hydrogen and helium (apart from a thin superficial layer) by nuclear burning. They self-sustain from gravitational collapse with the degeneracy pressure of a fermion gas of electron. This sets an upper-limit on the possible mass of these object called Chandrasekhar mass. This is $\sim 1.4M_{\odot}$ for a C-O white-dwarf.

It is largely accepted that SNe Ia originate from the thermonuclear disruption of a white dwarf, but apart from that, the exact nature of the progenitor and the mechanism of the explosion are heavily discussed and uncertain (Hillebrandt et al., 2013). Many of the observational characteristics of SN Ia can be naturally explained. Their uniformity, the small mass of the ejecta, the absence of helium and hydrogen, the absence of a compact remnant and of a bright progenitor. However, white dwarfs are thought to be stable and slowly cool without without exploding. Hence, SNe Ia are the result of the disruption of a white dwarf in a binary system, where the companion star is needed to trigger the explosion. However, the nature of the companion star, whether it is another white dwarf (Iben & Tutukov, 1984; Webbink, 1984) or a non-degenerate companion (Whelan & Iben, 1973; Nomoto, 1982) is still an open question. In these two scenarios, models differ from each other by the amount of mass gathered by the primary white dwarf at the time of explosion, the mode of thermonuclear combustion or the ignition mechanism (Hillebrandt & Niemeyer, 2000; Wang & Han, 2012; Hillebrandt et al., 2013).

1.3 Observational Efforts

Also fueled by the important use of SN Ia for cosmology, SN Ia are not only intensely studied theoretically, but are also heavily observed. There are many ongoing observational campaigns like the Nearby Supernova Factory (SNfactory, Aldering et al., 2002), the Palomar Transient Factory (PTF, Rau et al., 2009) or the Public ESO Spectroscopic Survey of Transient Objects¹ (PESSTO, e.g. Maund et al., 2013) and a large number

¹<http://www.pessto.org>

of SN spectra collected by the CfA Supernova Data Archive (Blondin et al., 2012), the CSP sample (Folatelli et al., 2013), the Berkeley sample (Silverman et al., 2012), and SN catalogs as SUSPECT² and WISEREP (Yaron & Gal-Yam, 2012). The number of well-observed SNe Ia has become large enough to allow for a quantitative statistical analysis of their spectral and photometrical diversity. Likewise, the complexity and diversity of synthetic spectra have increased (Hillebrandt et al., 2013), for the first time producing enough synthetic data to allow a coherent comparison between theoretical predictions and observations, although such a deep investigation is still to be reported.

1.4 Our current work: Statistical Methods for the Study of Type Ia Supernovae

In order to explore the potential of detailed high-quality observations and extended modelling, we aim at developing an enhanced framework where all information stored in a particular data set can be automatically used to characterize a given synthetic spectrum. This new metric space is constructed using an extended version of the Principal Component Analysis (PCA) method. PCA has been successfully used to classify QSO spectra (e.g. Boroson & Green, 1992; Francis et al., 1992; Yip et al., 2004; Suzuki, 2006), and it has become a standard technique in that field. It is also widely used for studying galaxy spectra (e.g. Connolly et al., 1995) and stellar spectra (e.g. Whitney, 1983; Bailer-Jones et al., 1998). A non-linear extension of PCA has also been used to photometrically classify SNe, in anticipation of the comparatively scarce spectroscopic resources to be faced by future cosmological surveys (Ishida & de Souza, 2013). Standard linear PCA was applied to SN Ia spectra recently by James et al. (2006) and Cormier & Davis (2011). Both papers concluded that PCA can be useful to study the diversity among SN spectra once larger samples become available.

In what follows, we will use an Expectation Maximization PCA (EMPCA) algorithm as implemented by Bailey (2012), which is capable of handling missing data and measurement uncertainties. The potential of information extraction enclosed in EMPCA is enhanced by pre-processing filtering and derivative routines, as well as by the use of complete spectral sequences in the construction of the initial data matrix. Once a stable PC space is obtained, we use Partial Least Square (PLS) analysis to demonstrate that the information it contains is not restricted to spectral indicators (velocities and pseudo-equivalent widths) but, as expected, it also correlates with photometric features. The outcomes from this analysis, applied to data from SNfactory, enable the construction of a metric space where any given synthetic spectrum can be projected and automatically confronted with real data. Systematic comparisons of models with observation are explored (e.g. Diemer et al., 2013, comparing light curves). Here we approach the problem from a new observation-driven perspective and we focus on spectral series.

²<http://www.nhn.ou.edu/~suspect>

This thesis is organized as follows: In section 2 we present details about all pre-processing techniques and statistical methods used in building our framework. The method is presented as a general data analysis technique which allows its application to any set of spectral sequences. The connection with SN Ia data is presented in the following sections. Chapter 3 describes the SNfactory data set, a data set built with public data, and the additional spectroscopic and photometric features to be investigated through the PLS algorithm. Results from the EMPCA analysis based on SNfactory data are presented in chapter 4, and in section 4.4 we present the EMPCA analysis of a collection of publicly available spectra. Chapter 5 presents the independently measured SNe Ia features investigated in this work and the corresponding results from PLS are shown in section 5.1. Chapter 6 studies in detail the major explosion mechanisms that are considered as explanations of SN Ia. Chapter 7 extends the ideas of PLS regression to predict light curves and color curves of SNe Ia. This allows to study extinction by dust using an innovative point of view. The characteristics of extragalactic dust is one of the open problem in astrophysics, and studies of SN Ia have been giving contrasting results. Finally, our conclusions are delineated in section 8.

Chapter 2

The Method¹

2.1 Filtering the Spectra

Before attempting any process of information extraction from spectral data, one must take into account the high impact of random noise originating in the observational process. Spectra are affected by noise arising from photon statistics, detectors, and calibration. Ideally, we would like to extract the features filtering the noise without degrading the signal.

2.1.1 Weighted Savitzky-Golay Filter

The Savitzky-Golay (SG) filter (Savitzky & Golay, 1964) is sometimes used to tackle this issue (Bailey, 2012; Poznanski et al., 2010; Hügelmeyer et al., 2007). It uses a least-square approach to fit a polynomial to neighbouring points within a fixed window around each wavelength. In comparison with other smoothing methods (e.g. simply re-sampling the data in larger wavelength bins), the SG filter, with an appropriate choice of parameters, is more successful in preserving the shape of the peaks and valleys, even for weak spectral features. The procedure is effective especially if the line broadening is significantly larger than the size of the wavelength bin as is the case here. Ideally, the smoothing window (polynomial degree) should be chosen such that it is not too small (large) to fail to filter the noise at the same time that it is not too large (small) so weak features are completely wiped away.

In this work, we wish not only to properly smooth a noisy spectrum, but we look for a procedure that takes into account the uncertainties associated with each measurement. Moreover, we should be able to calculate all the coefficients of the polynomial fit as well as their covariance matrix. In order to fulfil these requirements, we substituted the least square polynomial fit in the standard SG filter, by a weighted least square routine², where

¹Part of this chapter is published in Sasdelli et al. (2015)

²<http://docs.scipy.org/doc/numpy/reference/generated/numpy.polyfit.html>

the quantity to be minimized is given by

$$S = \sum_{i=1}^N [w_i (F_{\lambda_i}^{\text{obs}} - g_M(\lambda_i, \boldsymbol{\beta}))]^2. \quad (2.1)$$

Here, N is the number of data points included in a fixed window, $F_{\lambda_i}^{\text{obs}}$ is the observed flux at wavelength λ_i , g_M is the polynomial of degree M , $\boldsymbol{\beta}$ is the vector of scalar coefficients of g and w_i is the weight assigned to $F_{\lambda_i}^{\text{obs}}$. The algorithm returns the best fit values and covariance matrix for $\boldsymbol{\beta}$ at each wavelength. The width of the window is kept constant in $\log(\lambda)$, which corresponds to a constant velocity broadening to allow for a reasonable smoothing up to the minimum line broadening of the lines. Other types of smoothing techniques significantly improve the results of our analysis. This matter is investigated in some detail in the next subsection.

2.1.2 Improving the Filtering

The filtering is one of the crucial parts of our approach and in the literature there are more advanced techniques to filter the noise than a simple SG-filter (Savitzky & Golay, 1964). A well known way to improve low band-pass filters is to iterate the filtering a number of times (Kaiser & Hamming, 1977). To obtain a similar minimum band-pass the parameters of each iteration have to be looser. With this filtering we improved the rejection of the noise and at the same time obtained sharper spectral features to be fed into the PCA algorithm. This is a very simple approach and it proved to be very effective on spectra. With a larger number of iterations the window of filtering needs to be reduced in order to have a comparable effective window.

Once the impact of noise is reduced, we proceed to the construction of a framework capable of extracting information from a large data set, while minimizing the number of random variables to be dealt with.

2.2 Expectation Maximization Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction method used to describe an initially multivariate data set using a smaller number of uncorrelated parameters (principal components — PC). It transforms the original high-dimensional space through a rotation of its axes. The first new axis (or PC) is aligned with the direction of largest variance in the data. The second PC should also maximize the variance, subject to being orthogonal to the first, and so on. Mathematically, these directions can be more easily determined through the covariance matrix,

$$\Sigma_{ii'} = \frac{\sum_{k=1}^{k=N} (X_i^k - \bar{X}_i)(X_{i'}^k - \bar{X}_{i'})}{N}, \quad (2.2)$$

where \overline{X}_i is the mean of all fluxes measured at wavelength i and N is the total number of objects (for a complete review, see Jolliffe, 2002). Hereafter, we will always refer to the initial data as the mean subtracted terms in Eq. 2.2 (the centralized version of all points in the initial data set).

Once Σ is diagonalized the PCs are given by its eigenvectors, with the first PC corresponding to the one with the largest associated eigenvalue, and so on. We are now able to fairly reconstruct a given spectrum from the original data set using only M PCs ($M \ll N$),

$$\mathbf{F}_{\text{rec}} \approx \overline{\mathbf{X}} + \sum_{j=1}^M c_j \mathbf{P}_j, \quad (2.3)$$

with $\overline{\mathbf{X}}$ representing the mean of all spectra, \mathbf{P}_j the j -th PC and c_j the j -th scalar whose values must be determined from fitting \mathbf{F}_{rec} to the measured flux. Geometrically, c_j represents the projection of the measured spectrum on \mathbf{P}_j . PCA is just a basis change. Using all the N components the reconstruction becomes identical to the original data. The point is that the new basis captures a large fraction of the variance in a small number of components (M). For the purpose of this work, the determination of the ‘‘optimal’’ M is not a crucial point. A deeper discussion and other important applications of PCA for reconstruction in astronomy can be found in Ishida & de Souza (2011); Ishida et al. (2011); Benitez-Herrera et al. (2012, 2013) and references therein.

If a particular measurement is missing, or is not reliable enough to be considered on the same basis as the other more accurate ones, it is possible to reconstruct it from the nearest ones. Here we chose a different approach, taking advantage of a technique able to deal with missing elements in the initial data matrix: an expectation maximisation algorithm of PCA, first developed by Roweis (1998). We use an extended version of it, which can deal with non-uniform errors in the known components (Dempster et al., 1977; Bailey, 2012).

Reversing the line of thought which leads us to equation 2.3, we can think of the PCs as the vectors which minimize $\chi^2 = \sum_{k=1}^N [\mathbf{X}^k - \mathbf{F}_{\text{rec}}]^2$. In the presence of measurement errors, one can add a $k \times i$ weight matrix, \mathbf{W} , which controls the degree of influence of each flux measurement (for object k at wavelength i) in the determination of the components,

$$\chi^2 = \sum_{k=1}^N \mathbf{W}^k [\mathbf{X}^k - \mathbf{F}_{\text{rec}}]^2. \quad (2.4)$$

The above expression presents the challenge of diagonalizing a possibly very large matrix with a non-negligible number of null elements. Within EMPCA, this problem is tackled through the use of an *Expectation Maximization* algorithm (explained in detail in section 5.3 of Bailey, 2012).

This method allows us to perform PCA on real data by giving higher weight to points with lower noise. Moreover, missing components in the input data are handled easily by assigning them a weight equal to zero. Using the SG filter and EMPCA, we are able to

Algorithm 1 Expectation Maximization algorithm

1. $\mathbf{V} \leftarrow$ random orthonormal basis of dimension $i \times M$
 2. repeat until convergence (i.e. the basis \mathbf{V} does not vary significantly with new iterations):
 - (a) calculate the projections of all spectra on the basis \mathbf{V} (E-step)
 - (b) using these coefficient values, find a new estimate of the basis \mathbf{V} which minimizes equation 2.4 (M-step)
 - (c) normalize the columns of \mathbf{V} to unit length
 3. return \mathbf{V} as the EMPCA calculation of the first M eigenvectors of the basis \mathbf{P}
-

translate a set of spectra from wavelength to PC parameter space, with the SG filtering being crucial to ensure stability of the EMPCA results. In the absence of such filtering the EMPCA procedure does not converge to a stable solution.

2.3 Error Budget

The propagation of the errors from the spectra to the projections is not included in the EMPCA framework. For standard PCA, the error in the determination of each eigenvector is inversely proportional to the corresponding eigenvalue (Jolliffe, 2002). In EMPCA however, we need to deal with three main sources of error when analysing the geometrical distribution of our data in PC space. First, the iterative nature of the EM algorithm prevents us from obtaining the complete eigensystem and leads to uncertainty in the determination of the PC themselves. Beyond that, in the presence of missing data, computing the eigenvalues can be complicated, as it would require defining the total covariance based on an incomplete data sample. Second, once the PCs are given, we need to tackle properly the potential variance in their projections due to missing elements in the data vectors. Third, there is the contribution to the variance of the projections due to noise.

The determination of the PCs in EMPCA starts with a random first guess. It rapidly converges to an approximate final solution, but continues to fluctuate weakly even after many more iterations. The output PC vectors also vary slightly for different choices of the initial random seed. Despite the small influence of these features on the overall behaviour of our results, we took them into account by running the EMPCA algorithm for 100 different seeds during 500 iterations each. The resulting sets of vectors were then used to estimate the uncertainty in the projections in PCs space. A small value of these variances

can be interpreted as evidence that the input data quality is high enough to allow a stable determination of the PCs.

The errors in the projections due to missing measurements in the projected vector were calculated assuming that the eigenvectors are well determined, using the approach of Nelson et al. (2006). The propagation of the errors is due to the operation of projecting a non-complete spectrum on the PC space. The approach involves the inversion of submatrices of the covariance matrix, whose dimension is much larger than the sample size. An estimate of this matrix was achieved by completing the observed data with the PCA reconstructions. Then, we computed the estimator for the covariance of the completed data as described by Ledoit & Wolf (2004). With the covariance matrix and the eigenvectors we computed the error in the projection due to missing data for each object, as described by Nelson (2002), section 3.2.1 and Nelson et al. (2006).

The errors on the projections due to measurement noise were computed using a Monte Carlo approach. Each spectrum is submitted to the SG filter and a random noise based on the original error amplitude is added to the smoothed spectrum. The new noisified spectrum is again submitted to the filtering process and its corresponding projection in PC space are computed. The procedure is repeated 25 times. This allowed us to assess, in an empirical approach, the variance in the projections due to different magnitudes and covariances among the measurement errors.

2.4 Optimizing Information Extraction

After the smoothing described in section 2.1, we are left with a well behaved representation of the measured spectra. Mathematically, this would be enough to feed the EMPCA algorithm and perform the exercise of looking for patterns/subgroups in PCs space (e.g. Whitney, 1983; Francis et al., 1992; Connolly et al., 1995). However, astronomical spectra commonly also present uncertainties in large wavelength modes due to reddening, calibration problems, and on the absolute flux itself due to poor estimates of the distance of nearby galaxies. They can also present uncertainties on small wavelength modes due to CCD fringing at higher wavelengths, discontinuities in the overlapping region between spectra obtained with different spectrographs, or poor subtraction of telluric lines. In this context, our goal is to optimize the power of information extraction as much as possible, getting rid of any recognizable additional noise and enhancing intrinsic spectral features which we know to be relevant for individual object characterization.

2.4.1 Derivative Spectroscopy

Although we are aware that it is not possible to completely remove the effect of extinction in measured spectra, we can make it easier to handle by, first, using the logarithm of the flux as our initial data. As an example, consider a general reddening law:

$$F_{\log} = \log_{10} F_{\lambda}^{\text{obs}} = \log_{10} F_{\lambda}^{\text{intr}} - 0.4 \frac{A_{\lambda}}{A_V} R_V E_{B-V}, \quad (2.5)$$

where F_{λ}^{obs} , $F_{\lambda}^{\text{intr}}$ and A_{λ} are the observed flux, intrinsic flux and extinction at wavelength λ , respectively. A_V represents the extinction in the V -band and $R_V = A_V/E_{B-V}$, and A_{λ}/A_V is traditionally used to characterize the dust responsible for the extinction. From this expression we realize that in terms of F_{\log} , reddening becomes a linear relation in the extinction parameter E_{B-V} . Moreover, two objects following the same extinction law but subjected to different amounts of reddening will differ only by a multiplicative constant.

We would also like to take full advantage of the PCA dimensionality reduction power by equally weighting the information contained in weak/strong spectral lines. The presence of strong lines naturally dominates the variance (and consequently all results from PCA) of any given spectra data set. They are crucial for the initial classification, but in a second order analysis they may obscure important information contained in weak spectral features, which are more sensible to the conditions of the material because usually they are not saturated. It is important to emphasize that PCA itself is an excellent framework to study a “forest” of weak lines since this kind of study demands the parallel analysis of many of them.

We independently rediscovered a technique used in chemistry since Morrey (1968), which consists of beginning the analysis from the derivative of each spectrum over the wavelength, which in our case translates to $\partial F_{\log}/\partial\lambda$, hereafter dF_{\log} . This approach presents a few important improvements over the standard scenario for spectra analysis with PCA:

- Weak lines are emphasized. PCA on the derivative accounts for variance in the slope instead of variance in the flux, which also enhances the importance of the velocity of lines.
- It does not depend on errors in distances or on small calibration errors of each spectrum, since a change in any of these adds a constant to F_{\log} but leaves its derivative unchanged.
- It is only mildly dependent on reddening and large but smooth calibration errors, since these add a function to F_{\log} which is weakly dependent on wavelength (section 4).

2.4.2 Complete Spectral Sequences

The procedure described up to now can be applied to any data set composed of at least one spectrum per object. In a few cases however, mainly concerning transients, a specific data set will contain a sequence of spectra for each of its objects, taken at different epochs. When this is the case, we could, in principle, restrict ourselves to a single important epoch which would mean wasting a large part of the available information. Such a time-focused analysis would have no means of recognizing distinct evolutionary tracks for two

objects which happen to present similar features at the chosen epoch. Similarly, it would overestimate the distinction among two sources sharing almost identical spectral time evolution if they are submitted to external effects which are mainly detected at the time of observation (such as noise, or bad atmospheric subtraction).

Alternatively, one could compare results from the analysis of spectra taken at different epochs and follow the different PC space configurations over time. Although this naively seems a good option, it poses some difficult technical problems. Comparing PCA results from two different matrices would require spectra for all sources taken at exactly the same epochs (or within the same epoch bin) in order to have enough statistics to justify a PCA in each one of them. As this is not the case for current data sets, we chose to analyse all available spectra in a single PC space by concatenating subsequent spectra in each line of the initial data matrix. In this context, if one particular object is missing one spectrum the corresponding slots for those measurements are assigned a zero weight, and the EMPCA algorithm still uses the available data in the determination of the complete PC space.

2.5 The Partial Least Square Analysis

We now have a few techniques enabling us to translate the measured spectra from wavelength into PC parameter space. This new optimized space summarizes the information contained in the original data, grouping objects similar to each other and providing a low-dimensional basis from which we can reconstruct the main aspects of observed spectra. However, given that the PC space represents the essential information contained in each spectral sequence, it should be possible to obtain additional information from the PCs. It is reasonable to assume the existence of correlations between physical characteristics and a space that represents all spectral features, and in such case we would be able to associate known physical characteristics to the parameters found with EMPCA. In this context, we could easily recognize a missing or unexpected element in synthetic spectra. In this sub-section we show that the PLS analysis is suited for this task.

The Partial Least Squares analysis (PLS, also known as Projection to Latent Structures) is a technique used to find hidden relations between two groups of variables, originally developed by Wold (1982); Wold et al. (1984). The underlying hypothesis behind PLS is that all observed data are generated by a small number of latent variables, not directly observed or measured. It searches for traces of these latent structures which may be present in different parameter spaces.

We can roughly think of PLS as a combined principal component search. Suppose we have two independent sets of variables, $\{\mathcal{X}, \mathcal{Y}\}$, which result from measurements performed on the same objects. For example, \mathcal{X} can be a set of spectra and \mathcal{Y} the set of independently measured photometric properties of the same objects. If we apply PCA to each one of these sets individually, we would obtain two distinct groups of PCs and their corresponding data projections, but the PCs of \mathcal{X} would bare no information about the PCs, or projections, of \mathcal{Y} , and vice versa. The goal of PLS is to determine directions within \mathcal{X} and \mathcal{Y} that

maximize the covariance between their projected data. Once the directions are known, from measurements of a new object in \mathcal{X} we can estimate its projections and predict the values for variables in \mathcal{Y} .

In this work, we look for relations between a 1-dimensional parameter space \mathcal{Y} and the M -dimensional PC space coming from EMPCA. Mathematically, we are searching for the direction \mathbf{e} ($\sum_i e_i^2 = 1$) that maximizes

$$Cov(\mathbf{e}X, Y) = \frac{\sum_{k=1}^N (Y^k - \bar{Y}) \sum_i (X_i^k - \bar{X}_i) e_i}{N}, \quad (2.6)$$

where N is the number of objects and \bar{X}_j and \bar{Y} are means:

$$\bar{X}_j = \frac{\sum_{k=1}^N X_j^k}{N}, \quad \bar{Y} = \frac{\sum_{k=1}^N Y^k}{N}.$$

The corresponding correlation is the covariance weighted by the variances:

$$Corr(e_i) = \frac{Cov(e_i)}{\sigma(\sum_i X_i e_i) \sigma(Y)},$$

where

$$\begin{aligned} [\sigma(\sum_i X_i e_i)]^2 &= \frac{\sum_{k=1}^N (\sum_i (X_i^k - \bar{X}_i) e_i)^2}{N}, \\ [\sigma(Y)]^2 &= \frac{\sum_{k=1}^N (Y^k - \bar{Y})^2}{N}. \end{aligned}$$

PLS does not maximize the correlation, as the standard least square linear regression does, because that would assign the same weight to all directions in \mathcal{X} . Instead, it maximizes the covariance, which gives more weight to directions in \mathcal{X} with larger variance (first PCs) and avoids overfitting problems. In this work, we use the PLS algorithm as implemented by the *scikit-learn* statistical suite (Pedregosa et al., 2011).

In principle it is possible to apply PLS before the PCA dimensionality reduction. However, given the large dimension of the original spectral sequence data, that would barely simplify the traditional approach. Moreover the EMPCA method allows us to deal with missing components and diverse weights, and consequently apply the method to many more spectra without discarding incomplete or significantly noisy data.

Chapter 3

Applications¹

In this section we apply the previously explained framework to SN Ia spectra from the SNfactory collaboration (section 3.1) and to a database of publicly available spectra (section 3.2).

3.1 The Nearby Supernova Factory

The SNfactory is an experiment carried out using the University of Hawaii 2.2m telescope, mounted at Mauna Kea. Its goal is to obtain a sample of well observed SNe Ia in order to improve the measurements of cosmological parameters (Aldering et al., 2002; Copin et al., 2006). Spectra are acquired through a two-channel Supernova Integral Field Spectrograph (SNIFS, Lantz et al., 2004), which simultaneously covers channels B (3200-5200 Å) and R (5100-10000 Å). Discovery is largely automated using images from the JPL's Near Earth Asteroid Tracker (NEAT) and from the QUasar Equatorial Survey Team with quantitative and traceable selection of SN candidates (Bailey et al., 2007). This removes biases induced by the reliance on existing galaxy catalogs. Precise calibration is carried out in order to ensure agreement with high-redshift SNe (Buton et al., 2013). The spectra are deredshifted with independently measured host galaxies redshifts (Childress et al., 2013). Telluric lines are properly removed and Milky Way extinction corrections are applied to all spectra (Schlegel et al., 1998). Each supernova is followed from before B -band maximum up to 40 – 45 days after peak, resulting in 10-15 flux-calibrated low resolution spectra for each object. Most of the observed SNe are at the low-redshift end of the smooth Hubble flow ($0.03 < z < 0.08$), which enables a small error in the determination of distance from peculiar velocities while still being well within the homologous expansion regime.

Consequently, SNfactory provides a considerably large and relatively homogeneous data set of SNe Ia spectra (151 SNe and 2323 spectra at the time of this analysis), ideal for the study of second-order features as the one proposed here. Since all spectra are

¹Part of this chapter is published in Sasdelli et al. (2015)

obtained with the same instrument, resolution and host subtraction routine, the data set is homogeneous enough to allow for intrinsic astrophysical features to produce non-negligible effects in PCA results. In what follows, we shall directly probe this argument by correlating the remaining variance in flux measurements with specific photometric and spectroscopic SN features (section 5.1).

It is important to emphasize that we chose the SNfactory as a first test of these tools because the outcome would certainly be less obvious if obtained from a less homogeneous sample. However, due to the incorporation of the SG filtering and the use of dF_{\log} , the method is flexible enough to be applied to a much more diverse SNe Ia data (e.g. Blondin et al., 2012; Silverman et al., 2012).

Data Treatment

The processed portion of the data set contains 151 SNe Ia (2323 spectra) from which we selected objects with at least one spectrum before, one after B -band maximum and a minimum of three observed epochs between -10 and $+10$ days around B -band maximum. The epoch B -band maximum was determined from the SALT2 light curve fitter (Guy et al., 2007) applied on magnitudes obtained from integrating BVR top-hat filters (Pereira et al., 2013). Applying such requirements reduced our sample to 119 SNe and 764 spectra. In the sample the decline in magnitudes in 15 days after the maximum is within 0.7 and 1.7, the $B - V$ color within -0.16 and 0.40. The redshifts of the SNe are within 0.007 and 0.12. Plots showing the distributions of these parameters in the SNfactory sample are shown by Chotard et al. (2011a) and by Childress et al. (2013).

Each spectrum was smoothed by means of the weighted SG filter (section 2.1), using a third order polynomial ($M=3$), and a 6000 km/s-wide window as filter parameters. Those values were chosen by visually inspecting some smoothed spectra; potentially, the choice of different values may provide further improvements. It is also important to emphasize that this filtering technique performs satisfactorily up to a certain threshold and starts to saturate for very noisy spectra. In this context, the uniformity and quality of SNfactory data allow us to apply the filtering without the need to discard spectra due to poor data quality.

Figure 3.1 is an example of how a measured spectrum is transformed at different stages of the pre-processing treatment. The top panel shows the measurements from the standard SNfactory reduction pipeline (blue-full) and the corresponding spectra after the SG filtering (green-dashed). The bottom panel presents the derivative of the same spectrum (red-dashed) and its centred counterpart (yellow-full), that is the difference between the derivative and the mean derivative. The mean derivative is the mean of all SNe. This last product of the spectra preparation was used as input to build the initial data matrix. In both panels, functions are artificially displaced along the vertical axis in order to improve clarity.

Once all preparations are done, each row in the data matrix is constructed by grouping into bins measurements taken in 2 days within each other. Thus, a SN with no missing

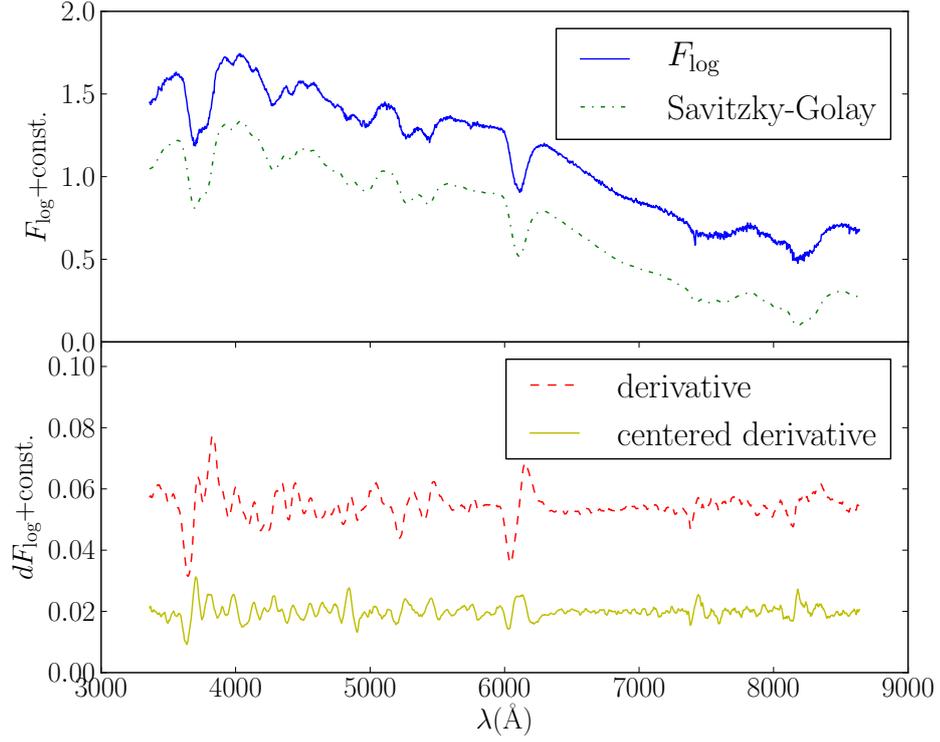


Figure 3.1: Multiple steps in data treatment. Both panels show data from SNF20080626-002, taken at -0.65 days relative to B -band maximum brightness. In each panel we artificially shifted the curves along the vertical axis for didactic reasons. **Top:** F_{\log} measurements before (blue-full) and after (green-dotted) going through the SG filtering. **Bottom:** dF_{\log} (red-dashed) and center derivative, $dF_{\log} - \overline{dF_{\log}}$ (yellow-full).

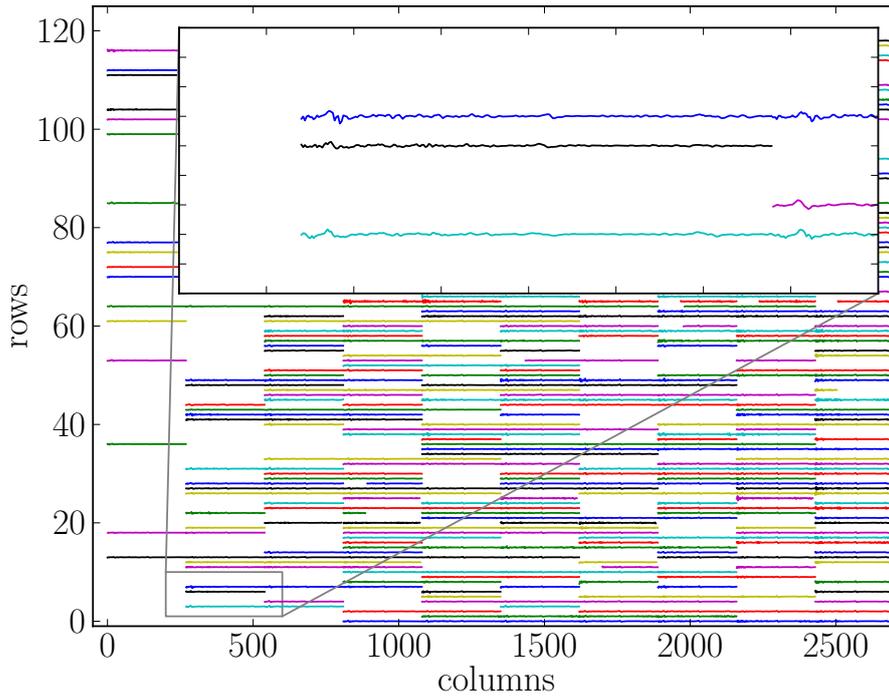


Figure 3.2: Representation of the input data matrix. Different rows correspond to different SNe. Each column shows centered dF_{\log} , from spectra collected between -10 and $+10$ days relative to B -band maximum brightness (from left to right), within each 2 day epoch window. Each curve runs over $3300\text{\AA} \leq \lambda \leq 9000\text{\AA}$, written in wavelength bins of 20\AA . The rectangle on top is the blown-up of the bottom-left corner of the data.

spectrum is represented by a row in the data matrix constructed from the concatenation of 10 spectra. The first was taken between -10 and -8 days, the second between -8 and -6 days, and so on. When a spectrum is missing, its corresponding matrix elements are left empty, and if more than one measurement exists within the same epoch bin, the mean spectrum is used as a representation of that SN in that bin. The choice of the parameters for the binning is inspired by the method of abundance tomography (Stehle et al., 2005; Mazzali et al., 2008). Using the SNfactory data, -10 days is as early as possible to have a rich sample. After $+10$ days the quality of the spectra generated by radiation transport codes not including forbidden line transitions starts to decrease (e.g. Sasdelli et al., 2014, for a study of the SN 1991T).

As with the SG filter parameters, the size of the epoch bin can be adapted according to the characteristics of each data set. For SNfactory, a two-day binning is a reasonable compromise, given that SN Ia spectra are quite homogeneous within this time frame and the data set is complete enough to provide a final matrix with more existing than missing spectra (in this configuration, we achieve 53% coverage). When transferring this procedure to another data set, one should keep in mind that an epoch bin should be small enough to guarantee that spectral variations between different objects within that bin are not due to time evolution. At the same time, the bins must be large enough to accommodate uncertainties in the determination of the epoch for each spectrum and allow a not too sparse initial data matrix.

Figure 3.2 illustrates the overall shape of the final data matrix. Each spectrum was sampled every 20\AA (wavelength gap between two columns for the same spectra). Our results show that this choice has negligible effects on the analysis and saves computational time.

In order to properly populate the weight matrix, errors coming from the flux measurements need to be propagated through the filtering process. Since the complete error covariance matrix of each spectrum is not used in the EMPCA code from Bailey (2012), we are computing only its diagonal terms. The weighted polynomial fit described in section 2.1 represents the smoothed spectrum at each wavelength as

$$F_{\lambda}^{\text{obs}} \rightarrow g_3(\lambda, \bar{\beta}) = \beta_0 + \beta_1(\lambda - \lambda_0) + \beta_2(\lambda - \lambda_0)^2 + \beta_3(\lambda - \lambda_0)^3, \quad (3.1)$$

where λ_0 is the central wavelength for each window. Given that each polynomial fit is used to determine the smoothed flux only at $\lambda = \lambda_0$, this implies that for each wavelength:

$$F_{\log} = \log_{10} F_{\lambda}^{\text{obs}} \Big|_{\lambda=\lambda_0} = \log_{10} \beta_0, \quad (3.2)$$

$$dF_{\log} = \frac{d \log_{10} F_{\lambda}^{\text{obs}}}{d\lambda} \Big|_{\lambda=\lambda_0} = \frac{\beta_1}{\beta_0 \ln 10}, \quad (3.3)$$

finally propagating the errors:

$$\delta F_{\log} = \frac{\delta\beta_0}{\beta_0 \ln 10}, \quad (3.4)$$

$$\delta dF_{\log} = \left| \frac{\beta_1}{\beta_0 \ln 10} \right| \sqrt{\frac{\delta\beta_0^2}{\beta_0^2} + \frac{\delta\beta_1^2}{\beta_1^2} - \frac{2\text{cov}(\beta_0, \beta_1)}{\beta_0\beta_1}}, \quad (3.5)$$

where $\delta\beta_i$ denotes the uncertainty associated with the determination of parameter β_i and the covariance between the first two parameters is represented by $\text{cov}(\beta_0, \beta_1)$. The weight matrix elements are then defined as $w_i = \delta F_{\log}^{-2}$ or $w_i = \delta dF_{\log}^{-2}$ for the logarithm and derivative cases, respectively.

There are a few supernovae within the SNfactory set whose errors are an order of magnitude smaller than the ones of the bulk of the data. This happens for bright SNe, where the number of counts is high and the Poisson error small. For example SN 2007le, being one of the nearest supernovae in the sample, has errors much smaller than most of the other objects. If the EMPCA is carried out with errors as they come out of the SG filter, it would overweight the two or three supernovae with the smallest errors and the first components would point in the direction of these few objects. This behaviour of EMPCA in the presence of few objects with a noise much lower than the rest of the sample is also highlighted by Bailey (2012, section 8.3). To overcome this problem, we artificially decreased the weight of 52 SNe (42% of the sample) in order to have no SN with a weight larger than 90 times the sum of the weights of the other objects. Results are not biased towards these objects and the PC space is stable as long as their number is kept between $\sim 25\%$ and $\sim 75\%$ of the total data set. We also performed the analysis without changing the initial weights, but removing the 8 SNe with lowest noise from the initial sample. The test returned the same results, demonstrating the low sensitivity of this procedure regarding the method used for down-weighting. Once the PC space is determined, the spectra are not downweighted to obtain the projections.

3.2 Publicly Available Spectra

In this section we make use of the set of techniques developed in in the previous chapters for the study of SN Ia spectral time series and photometry and we apply them to a database of public data. Publicly available SNe include a large number of very nearby ones with well observed low noise spectra. These publicly available data have a larger diversity in properties than what was observed by SNfactory and a large diversity in spectral properties is important for the comparison with models. In fact, the modelled explosion scenarios that we will discuss could be the explanation for normal SN Ia or for peculiar objects. Finally, among the public data there is a large number of well observed early spectra (earlier than a week before B -max). The early behaviour, on the other hand, is crucial to constrain models and, in addition the radiation transport codes used for the modeling work better for the early phases. Therefore, in order to construct a metric space for such an

analysis we collected a large sample of SN Ia spectra publicly available in literature. The sources are the CfA spectroscopic release (Blondin et al., 2012), the Berkeley Supernova Program (Silverman et al., 2012), the Carnegie Supernova Project (CSP, Folatelli et al., 2013). We are also using SN catalogs as SUSPECT² and WISEREP (Yaron & Gal-Yam, 2012). The spectra have been de-redshifted by using the heliocentric redshifts tabulated in Blondin et al. (2012) (CSP spectra are published in rest frame).

The photometry and the $B - V$ colors are collected from Hicken et al. (2009). They obtained them from light-curve fitting using MLCS2k2 (Jha et al., 2007). The observed photometry is obtained by filters in the observer frame (the observatory). The values in the object rest frame, typically redshifted due to the cosmic expansion, need to be obtained from the observed ones with interpolations that are called K-corrections (e.g Nugent et al., 2002). The photometry and colors were K-corrected, corrected for Milky Way extinction, and corrected for time dilation. The host-galaxy extinction was not removed. The CSP photometry comes from Stritzinger et al. (2011).

To estimate the distance of nearby galaxies from the measured redshift it is usually impossible to know the proper motion of the galaxy in question. But also our solar system has a non negligible proper motion, compared to the Cosmic Microwave Background (CMB). The proper motion of our solar system is well known and can be taken into account. This is usually done correcting the observed redshift by the projection of the proper motion of the Sun along the direction of the considered galaxy. This is called CMB centered redshift, i.e. the redshift that we would observe if we were in the CMB rest frame. The B -band photometry is transformed into absolute magnitudes using the CMB centered redshift measurements from Hicken et al. (2009). The error of the absolute magnitude is computed adding in quadrature an error due to the peculiar motion of the host galaxy. We assume a standard deviation of 500 km s^{-1} for this peculiar velocity (Hawkins et al., 2003).

Data Treatment of the Public Data

As for the SNfactory data set, the input matrix for the analysis is constructed treating every supernova as an individual observation and treating the spectra at different epochs as different observables. The input vectors of this matrix are constructed concatenating spectra at different epochs in the spectral series. The EMPCA code from Bailey (2012) deals automatically with missing epochs and/or missing wavelength ranges in the data. The derivative analysis frees us from the need of flux calibrated spectra with known distances, and the large number of well observed low noise nearby SNe allows us to increase the included range of epochs. We now include spectra from -12.5 days up to $+17.5$ days from B -maximum. The spectral coverage of CfA supernovae, by number the largest of the sample, is usually limited to below $\sim 7000\text{\AA}$. This restricts our current analysis to a spectral range between $\sim 3500\text{\AA}$ and $\sim 7000\text{\AA}$. This is not an issue, however, since all the

²<http://www.nhn.ou.edu/~suspect>

information red-ward of 7000\AA (mostly in the IR triplet of Ca II) is also present in the included wavelength range.

Chapter 4

Principal Components Interpretation and Metric Space Comparison¹

We present below, side by side, results from the application of the EMPCA to SNfactory data, with matrices built from F_{\log} and dF_{\log} (Figures 4.1 and 4.2, and Figures 4.3, 4.4 and 4.5). Hereafter, the PCs derived from a data matrix based on F_{\log} will be referred to as $PCi^{F_{\log}}$, with i denoting the PC number. Alternatively, PCs calculated from a matrix based on derivatives will be simply called PCi . This direct comparison allows the reader to clearly recognize the differences and advantages in using the derivatives, which is a crucial step for the subsequent PLS analysis presented in section 5.1.

4.1 Principal Components

Figures 4.1 and 4.2 show the behaviour (first panel) and the contribution to the reconstructed spectra (second to fourth panels) of the first three eigenvectors for analyses based on F_{\log} and dF_{\log} , respectively. In both figures, the first panel displays the functional form of the PCs themselves, while the remaining panels show the effect we can achieve in the final reconstruction by increasing the weight assigned to each PC within the boundaries allowed by the data. The reconstructions presented here are non-cumulative. In other words, the gray region in each panel represents features which arise when combining the mean spectrum with each PC separately. From this, we see that the first eigenvector computed from F_{\log} (Figure 4.1) leads to a slow variation with wavelength in the reconstructed result. Its influence can be easily associated with a constant that allows a rigid translation in flux, although it also carries some discrete wavelength dependent features.

¹Part of this chapter is published in Sasdelli et al. (2015)

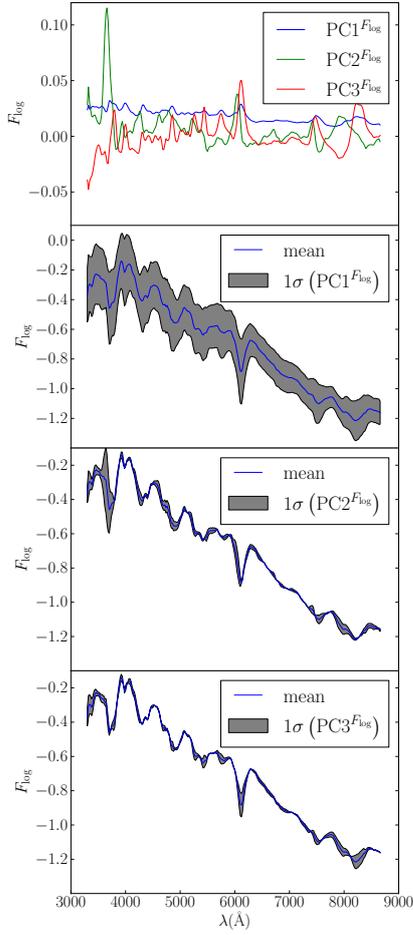


Figure 4.1: First panel shows the first three eigenvectors obtained from the analysis on F_{\log} . The second to fourth panels illustrate the main spectral features tracked by PC1, PC2 and PC3. All panels correspond to a spectrum taken between -6 and -4 days relative to B-band maximum. Blue lines denote the mean spectrum. Gray regions were obtained by reconstructing the spectrum with only 1 PC and varying the scalar coefficient within the 1σ range given by the data. The PC2 and PC3 bare similarities with the Si and Ca components found by Chotard (2011).

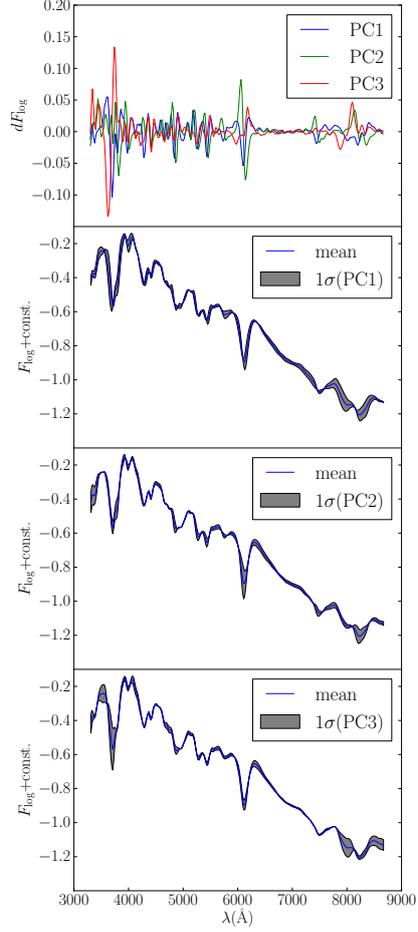


Figure 4.2: Same as Figure 4.1, but from a data matrix based on dF_{\log} .

Also, it clearly describes a much larger variance than the next two components (larger area covered by the gray region, second panel of Figure 4.1). The first PC is largely influenced by dust, with its long wavelength behaviour being consistent with a Cardelli et al. (1989) reddening law, the dust typical of the Milky Way. However, significant contributions to the flux and to the slope of this eigenvector due to absolute magnitude and intrinsic color variations are likely. The mixing of intrinsic and extrinsic properties is avoided by the PCA based on the derivative. For dF_{\log} (Figure 4.2, panels 2–4) one can notice that an important role is assigned to small scale variations. Moreover, the variance covered by the first PC is comparable to that of the others. This is a direct consequence of our choice of removing the overall flux information from the input data through the use of the derivative. In this analysis, the first three PCs show variations of pseudo-equivalent widths (pEW) and velocities of many lines, some of which are studied in more detail in section 5.1.

4.2 Metric Spaces

The projection of SNfactory data in a 2-dimensional PC space, obtained from F_{\log} , is displayed in Figure 4.3. Individual objects are coloured following the classification scheme defined by Wang et al. (2009), where high-velocity SNe are those whose velocity of the Si II 6355 Å is more than 3σ above its mean value. In what follows, we consider the mean $+3\sigma$ equal to 12200 km s^{-1} , as computed by Blondin et al. (2012). We also highlighted a few 91T-like SNe (red stars), following the classification used by Scalzo et al. (2012). 1999aa-like SNe are not highlighted as 91T-like. Crosses correspond to 1σ uncertainties due to random seed variation and ellipses represent the 1σ errors coming from missing data in the projected spectral sequence and measurement noise added in quadrature. After exploring a large range of the MC parameters, our results show that 25 realizations were more than enough to the secure stability of the error bars.

Figure 4.3 can be considered to be an alternative visualization of the same effect as presented in Figure 4.1: the first PC obviously contains a larger part of the total variance, and consequently the interpretation of the subsequent PCs is obscured. In this context, although we can identify a certain clustering of 91T-like SN in larger values of PC1, contamination is still significant, and an attempt to separate the set according to these features would certainly present important drawbacks. This high level of contamination is mainly due to reddening. This is shown clearly by the variation of the projections of SNF20080720-001 after a reddening correction of up to $E(B - V) = 0.4$ with a Cardelli et al. (1989) law (magenta line in Figure 4.3). This object has an observed $B - V$ color of ~ 0.4 , one of the reddest SNe in the SNfactory sample. Figure 4.4 shows the analogous situation for PC2 \times PC3 parameter space. The magenta line corresponds to the reddening effect still present in the second and third PC in flux space, showing that the PCA in fluxes is not able to isolate the effect of reddening in the first PC.

Figure 4.5 shows how this situation changes when the analysis is based on dF_{\log} . The

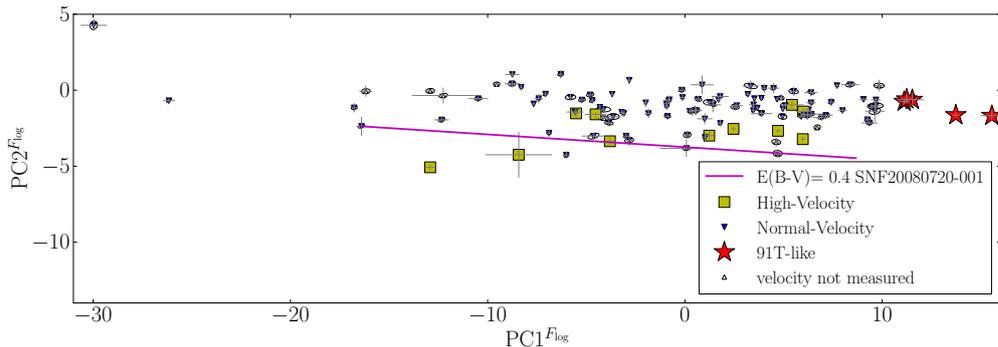
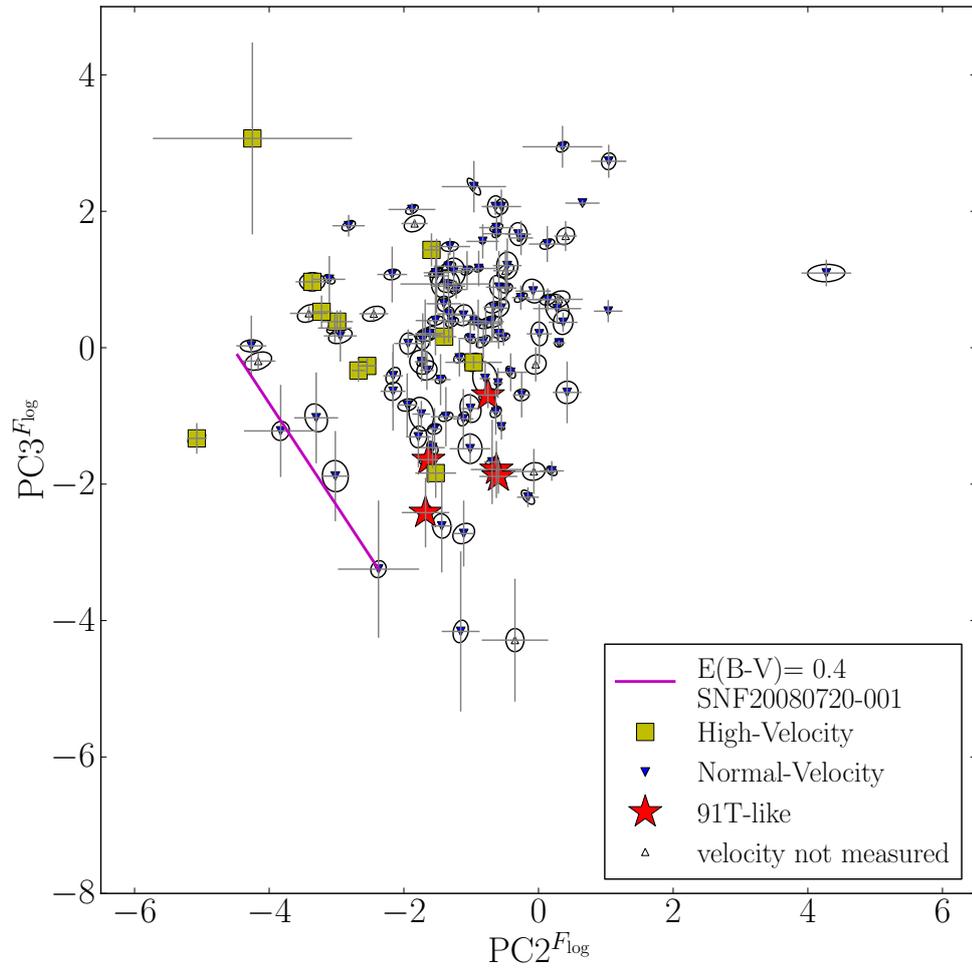


Figure 4.3: Projections of SNfactory data on the first two PCs for an analysis based on F_{\log} . Each point represents a supernova, colored according to the spectral classification of Wang et al. (2009). A few 91T-like SNe are also highlighted. The crosses correspond to 1σ errors coming from random seed variation and the ellipses denote 1σ uncertainties due to missing data and measurement noise. The magenta line shows the effect of reddening on the projection of the SN SNF20080720-001, which presents an observed $B - V$ color of ~ 0.4 mag.

crosses due to the instability of the EMPCA algorithm are completely negligible, the ellipses due to noise and missing components are large only for a few very noisy SNe. The slowly declining 91T-like SNe (red stars) are at the bottom edge of the diagram, clearly separated from the high-velocity ones on the right (yellow squares). The spectroscopically normal SNe (blue triangles) are spread throughout the parameter space, indicating a larger intrinsic variability between these objects. Visually inspecting spectra from the SNe in the upper-left corner, we also realize that this space is occupied by fast declining SNe with cooler spectra showing a lower ionization ratio. According to the projections in our metric space there are no clear separations that justify the definition of subclasses. SNe Ia, accordingly to spectral features, look like a continuous distribution of objects. In other words, there is no clear separation in velocity or EW of lines which justifies or objectively indicates a threshold for defining a subclass, although there are undoubtedly fundamental differences between objects in the extremes. For example, 91T-like SNe show a “bridge” of objects that connects them with the bulk of normal ones. The same is true for the ones with a high velocity of Si.

The marginal effect coming from reddening in this context is illustrated by the magenta line in Figure 4.5. As in Figure 4.3, it represents the translation in PC space experienced by SN SNF20080720-001 when a 0.4 mag reddening correction is applied. Comparing the magenta lines in both figures demonstrates the power of the derivative analysis in minimizing the effect of dust in the PC space. Although this is one of the most reddened

Figure 4.4: Same as Figure 4.3, $PC2^{F_{log}}$ and $PC3^{F_{log}}$.

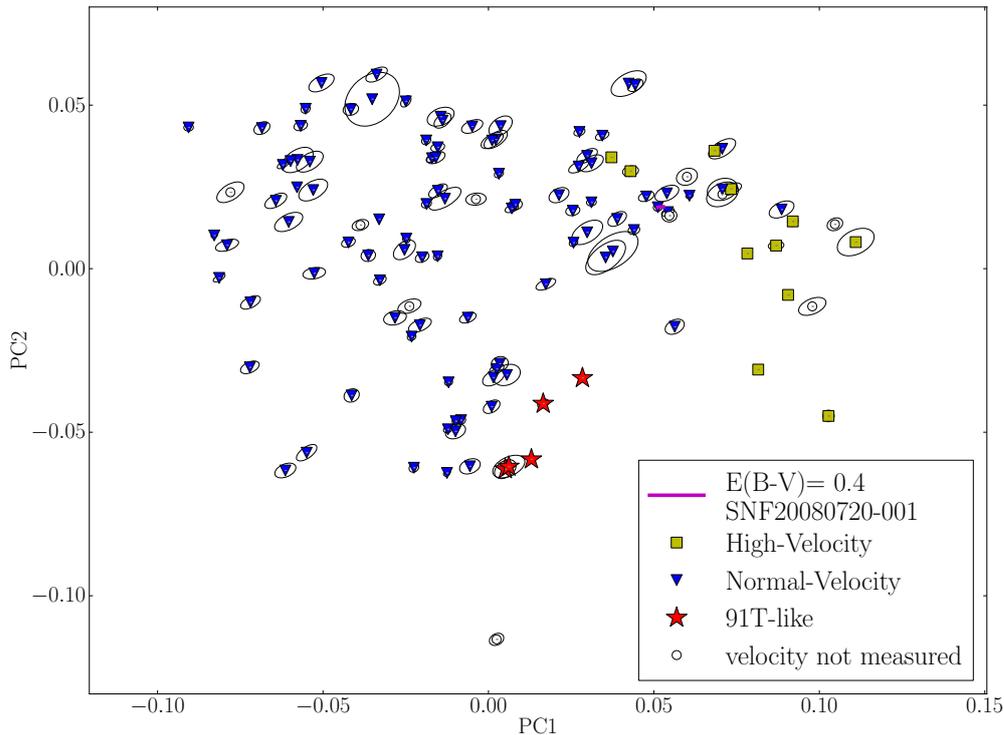


Figure 4.5: The figure shows the scatter plot of the first two components of the PCA analysis for dF_{\log} . The tiny magenta line shows the negligible effect of reddening on the values of the PCs. Three classes of SN Ia are highlighted: normals, 1991T-like, and those with high velocity of Si.

SNe, the change in the PCs is merely marginal. The same trend is observed for all the other objects in this sample.

It is important to keep in mind that this specific geometrical configuration in PC space will always be related to the sample of objects used to construct it, and it is not a “universal” space for SNe Ia. However, it is reasonable to expect that the addition of more high-quality data leads to an asymptotic PC space configuration which summarizes the similarities and differences within the SNe Ia sample used in its construction. Nevertheless, with the SNfactory data at hand, we are already able to demonstrate that the analysis is useful to look for correlations in the data, attack the problem of SN Ia spectra characterization and search for outliers.

Although this “universal” PC space is merely an asymptotic state, we can have a hint on how close it is to the ideal configuration. In other words, we can test the stability of a given PC space through the successive application of the EMPCA algorithm to different

subsets of the original data. This procedure is called Cross-Validation (CV) and it has been used in many fields where the configuration of a given method depends on the initial data set (Arlot & Celisse, 2010). Detailed results from a CV test are presented in Appendix A.1, and these demonstrate the stability of the space presented in Figure 4.5.

After analysing the first pair of PCs and confirming the stability of the PC space, we are left with an obvious question: how many PCs are necessary to describe the data set and throw away a substantial part of the noise? In a standard PCA the fraction of the total variance associated to each PC, or to a subset of them, can be estimated through the cumulative percentage of total variance (Jolliffe, 2002; Ishida & de Souza, 2011; Benitez-Herrera et al., 2013). Given that the eigenvalues associated with each eigenvector constitute a measurement of the data variance along that PC direction, this means that the ratio between the largest eigenvalue and the sum of all eigenvalues gives an estimative of the percentage of variance (or information) described by the first PC. However, in the EMPCA approach we do not have access to all eigenvalues at once, since the eigenvectors are calculated one at a time through the EM algorithm. Nevertheless, we do expect that only a handful of PCs will actually carry meaningful information and this hypothesis can be tested with a small sub-sample of them.

We used the EMPCA approach to calculate the first six PCs and their corresponding data set projections. From these, we determined the variance along each PC. By definition, the first PC contains a larger fraction of the total variance than any other PC, so we used it as a normalization factor. In this context, we can obtain an estimate of how much information is stored in a certain PC, in comparison to that in the first one.

In Figures 4.6 and 4.7 we show the variances normalized to the first component for the analysis on F_{\log} and dF_{\log} , respectively. Figure 4.6 shows the same result we have seen in Figures 4.1 and 4.3, with most of the information concentrated in $PC1^{F_{\log}}$. From a physical perspective, performing the analysis in this parameter space is challenging due to extinction and intrinsic luminosity variations. Extinction effects are present in all the principal components, making it difficult to disentangle two very different physical processes. For example, in this context two similar SNe subjected to different amounts of reddening would be distant from each other in the PC parameter space (as illustrated by the magenta line in Figures 4.3 and 4.4). On the other hand, when using dF_{\log} we concentrate the investigation on spectral features which are crucial to SNe Ia characterization and consequently a larger number of PCs are found to be significant. The derivative approach removes the effect of reddening, a physical process that causes a large amount of variance in the data, making it easier to train the PCA space. From Figure 4.7, it is clear that PC2 to PC5 carries at least 20% of the variance in PC1 each and the fractions stabilized for PC6. Thus, we conclude that 5 PCs are enough to describe most of the variance in SNfactory.

In Figure 4.8 some of the reconstructions are directly shown. We present the original spectra along with reconstructed ones using two and five PCs. The plot shows a few SNe at maximum for clarity, but this behaviour holds for all epochs between -10 to $+10$

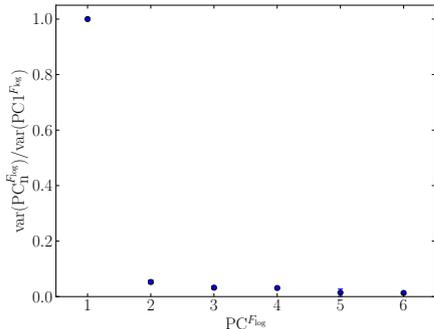


Figure 4.6: Distribution of variance among the first 6 PCs from F_{\log} data matrix. The variances are normalized to that of the first PC. The errorbars show the variability due to k-folding (Appendix A.1).

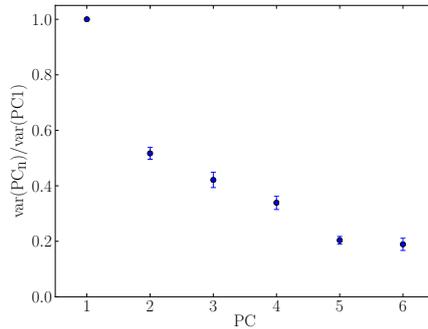


Figure 4.7: Same as Figure 4.6, but obtained from dF_{\log} data matrix.

days. Here, the consequences of our choices in focusing on intrinsic features are obvious. Although the overall spectral shape and most lines are very well recovered, the ratio of fluxes at long wavelengths (color) is not. This is welcome and expected because the derivative analysis does not give much weight to the mean slope of the spectra, making the analysis independent of individual SNe reddening and reshapes the observed spectra so to allow a fair comparison with synthetic models. The comparison in the derivative space is shown in Appendix A.3.

4.3 High Velocity Features

The first two PCs contain a large part of the spectral variance in the SNfactory data. This will be studied in detail in the subsequent sections. We highlight the significant role played by the third PC shown in Figure 4.2, which tracks the variation of the so called high velocity features (HVF) of Ca II H&K and infrared lines without particularly affecting the rest of the spectrum. This figure represents the eigenvectors in the epoch range between -6 and -4 days relative to B-maximum, since the high-velocity part of these lines usually disappears at later epochs. The third PC, by construction uncorrelated with the first two, seems to be mainly responsible for tracking variations of HVFs of Ca, thus, confirming that HVFs of Ca are a property of the outer layers of the ejecta and they are not correlated with the underlying structure (Mazzali et al., 2005). Such an effect can be achieved with an asymmetric/clumpy outer layer of the ejecta convolved with line-of-sight effects (Tanaka et al., 2006) and is a good indicator of the kind of astrophysical characteristics which can

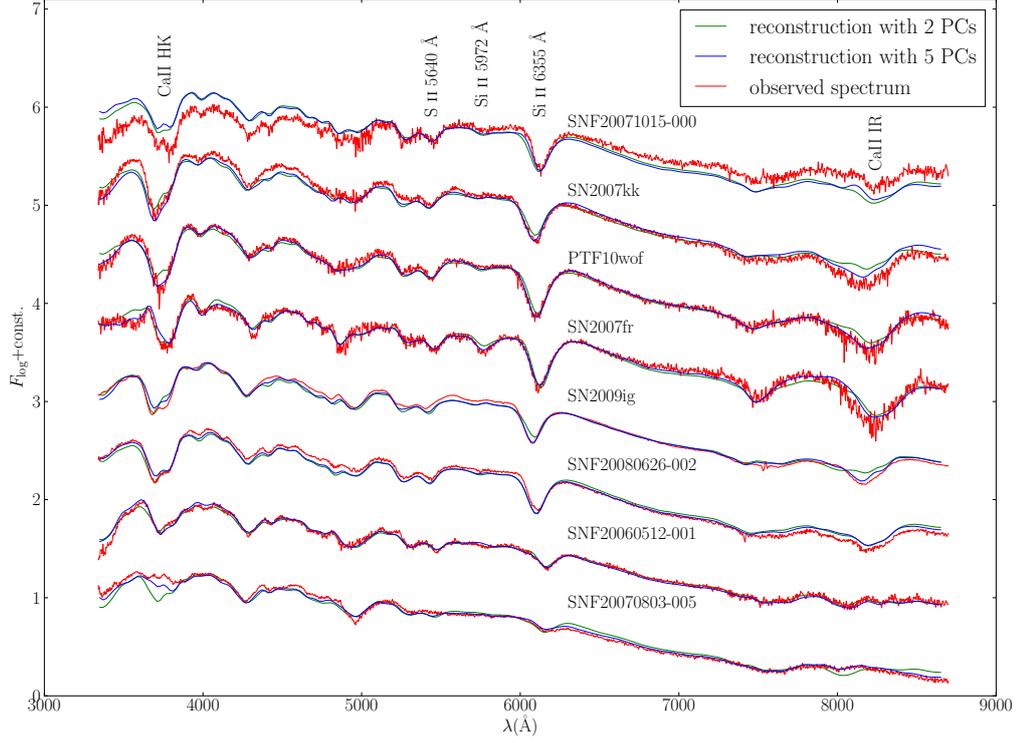


Figure 4.8: Comparison between the observed spectra without smoothing (red) and reconstructed spectra using 2 (green) and 5 (blue) PCs, in the dF_{\log} approach, for a few supernovae at B -band maximum light.

possibly be recognized also in synthetic spectra.

We want to investigate the behaviour of the HVF of the Ca II lines in some more detail. The high velocity component commonly appears in the features originating from two groups of lines: the Ca II H&K doublet in the blue part of the spectrum, and the Ca II IR triplet in the near-infrared. A subspace of our metric space encodes the information to describe these lines. We want to characterize this subspace, find its orientation and dimensionality. An easy way to do this is to compare the projections of the SNe in the metric space after removing the Ca lines from their spectra. That is, we treat the parts of the spectrum that show HVF of Ca as missing data. The new projections are different from the original ones by a vector that lies inside our subspace. Figure 4.9 shows how much the projections change in the first three components. The directions are strongly aligned, suggesting that they belong to a 1D subspace. To quantify this and to check the shape of this subspace of the 5 important dimensions we simply run PCA on these displacement vectors. The variances of this PCA are shown in Figure 4.10. The first component largely dominates over the others. This implies that the space is effectively 1D, and it means that the additional information needed to completely describe the HVF of Ca II is just one parameter. Since this direction is very different from the first two components for all the supernovae in the PCA space, it means that the physical mechanism behind the formation of the HVF of Ca II is very different from and uncorrelated with what drives the physics of the explosion. It is something that happens in the outermost layers of the ejecta only and that is not related with the mechanisms that produce the big diversity within the SN Ia class.

A possible explanation for the presence of the HVFs is interaction with circumstellar material. H in the circumstellar gas increases the electron density in the material. This is necessary to decrease the ionization state of the gas and to have a significant number of Ca II ions. Ca II has some of the strongest lines in the astrophysical context and shows up strongly in the spectra also in small quantities. In this picture the Ca lines would come from the metallicity of typical H rich material.

Another promising possibility to explain the HVFs of Ca is the detonation of He in an outer shell of the white dwarf. If this happens at the right density, it can produce an external layer mostly composed of Ca. The He detonation proceeds in the alpha process up to an element that is related to the density of the fuel. With the right density, about half of the He is converted into Ca without a significant amount of lighter or heavier elements of the alpha chain. He does not easily show up in the spectra, and it would not be easy to detect it. The variability of this feature is largely uncorrelated with the behaviour of the bulk of the explosion. This can be easily explained by differences in the mass of the He layer, if the burning is triggered at densities within the range that produce abundant Ca. It is not clear if an overabundance of Ca can explain the quick disappearance of the HVFs of the lines in the post-maximum spectra. A significant electron density due to He could lead to a quick change in the ionization state of Ca. In fact, the double detonation scenario predicts the detonation of a He layer on the surface of the white dwarf. In this

scenario the He detonation acts as the trigger of the explosion. This scenario is explained in detail in chapter 6.

4.4 Spectral Series and EMPCA on the Publicly Available Spectra

We applied our methodology on SN Ia spectra publicly available at the time of writing. The details of the database are described in section 3.2.

The metric space obtained from public data is very similar to the one obtained in our previous section from SNfactory data. The metric space resulting from PCA has also a low dimensionality. The output consists of just five significant components. Projecting the supernovae on the first three principal components (Figs. 4.11 and 4.12) shows the groups found in our previous section (Fig. 4.5). Normal SNe Ia are on the top-left side of the cloud of points in Fig. 4.11, supernovae with high velocity photospheric Si II 6355 Å have a large first component, 1991T-like events have negative second and third components. In the public sample we have also fainter supernovae. There is a significant number of 1991bg-like ones, characterized by fast declining light curves, low luminosity and low temperature of the spectra. At the bottom of Fig. 4.11, further apart than 1991T-like, there are a number of supernovae left unclassified by Blondin et al. (2012). Many of them are 2002cx-like. This is another elusive class of faint objects characterized by hot spectra and very low line velocities. These faint SNe Ia are not well represented in the SNfactory sample. The reconstruction of the original spectra are generally excellent. In Fig. 4.13 we show the comparison between the reconstruction and the spectra of SN 2008Z, a SN with a complete spectral coverage.

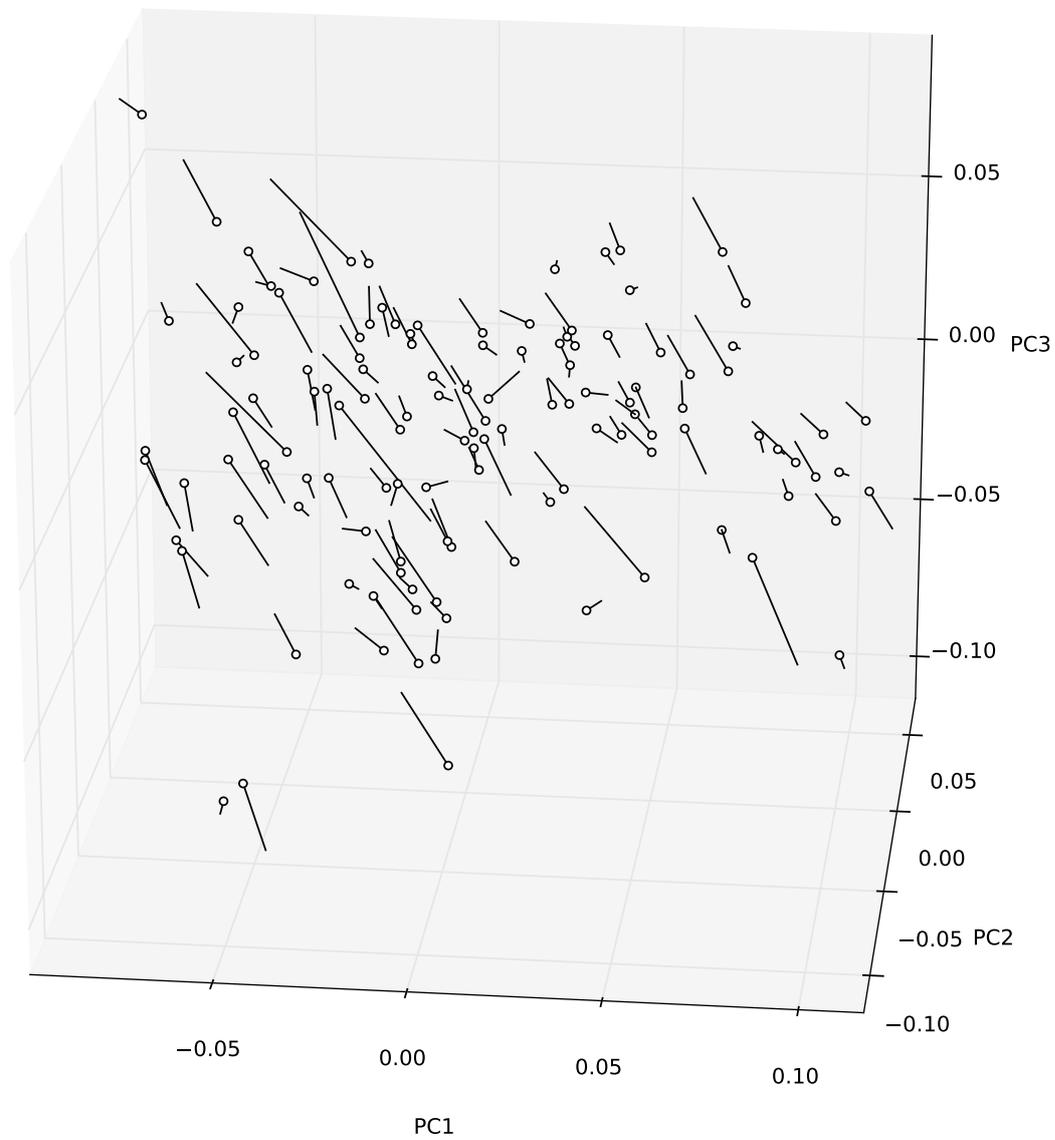


Figure 4.9: The figure shows the displacement in the projections due to the HVF of Ca II. The circles are the original projections. The tip of the line shows the projection without the information in the Ca lines.

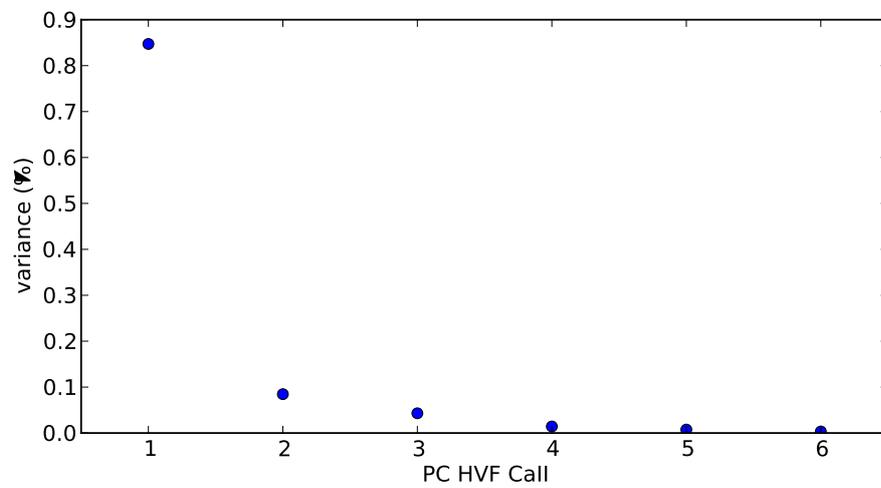


Figure 4.10: The figure shows the variances in the displacements shown in Figure 4.9. The first component is dominant. The subspace encompassed by these displacements is effectively 1D.

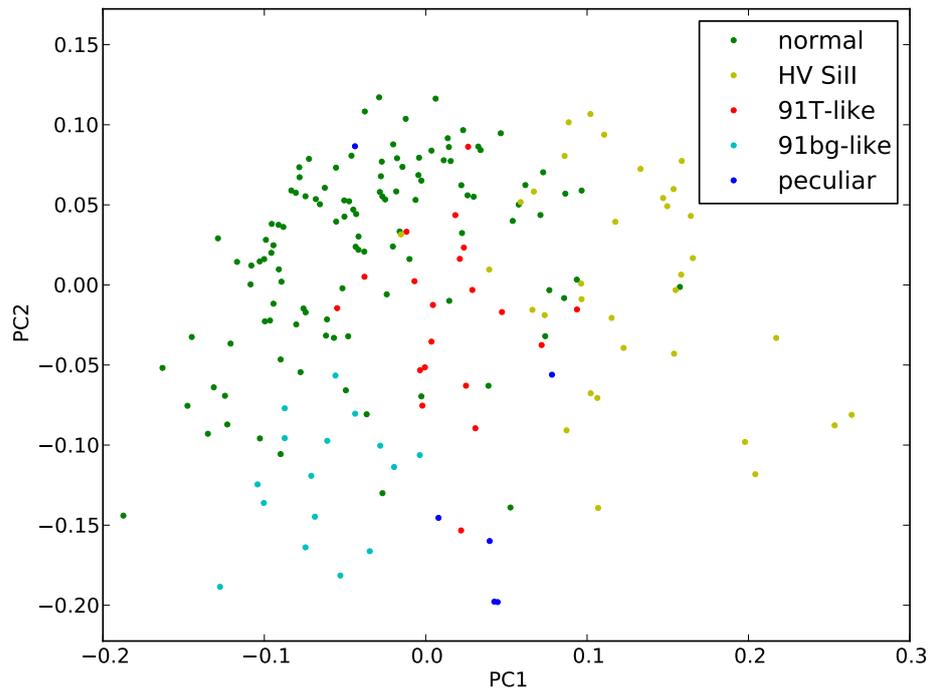


Figure 4.11: Plot of the first two principal components for the SN Ia of our public sample. The metric space obtained from the public SNe Ia clearly distinguishes the spectroscopic subtypes (color coded). The classification is due to Blondin et al. (2012).

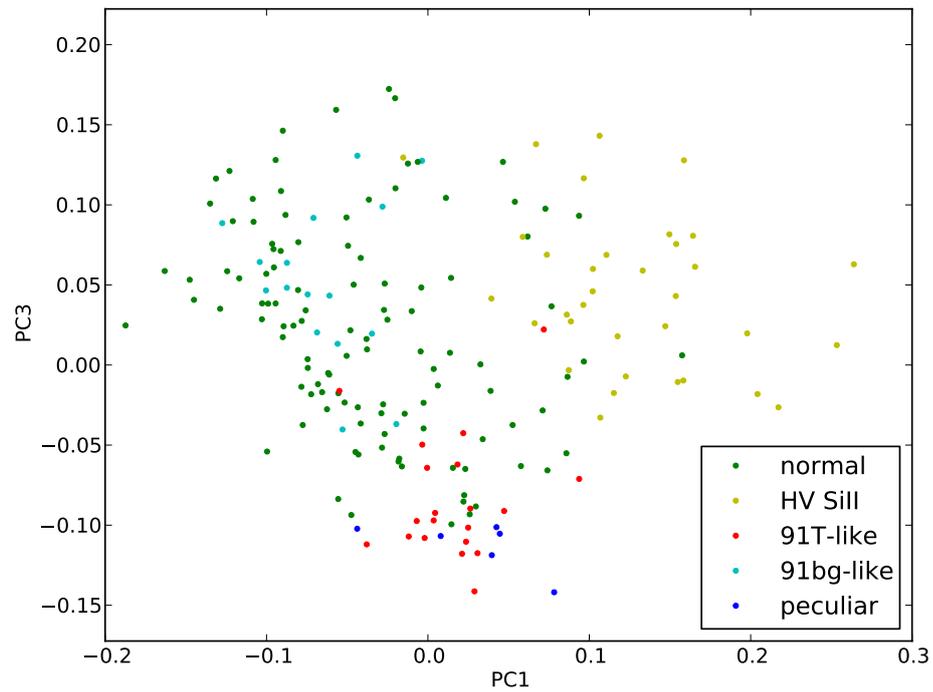


Figure 4.12: Plot of the first and the third principal components for the SN Ia of our public sample. This is a viewing angle on our metric space different from Fig. 4.11. The classification is due to Blondin et al. (2012).

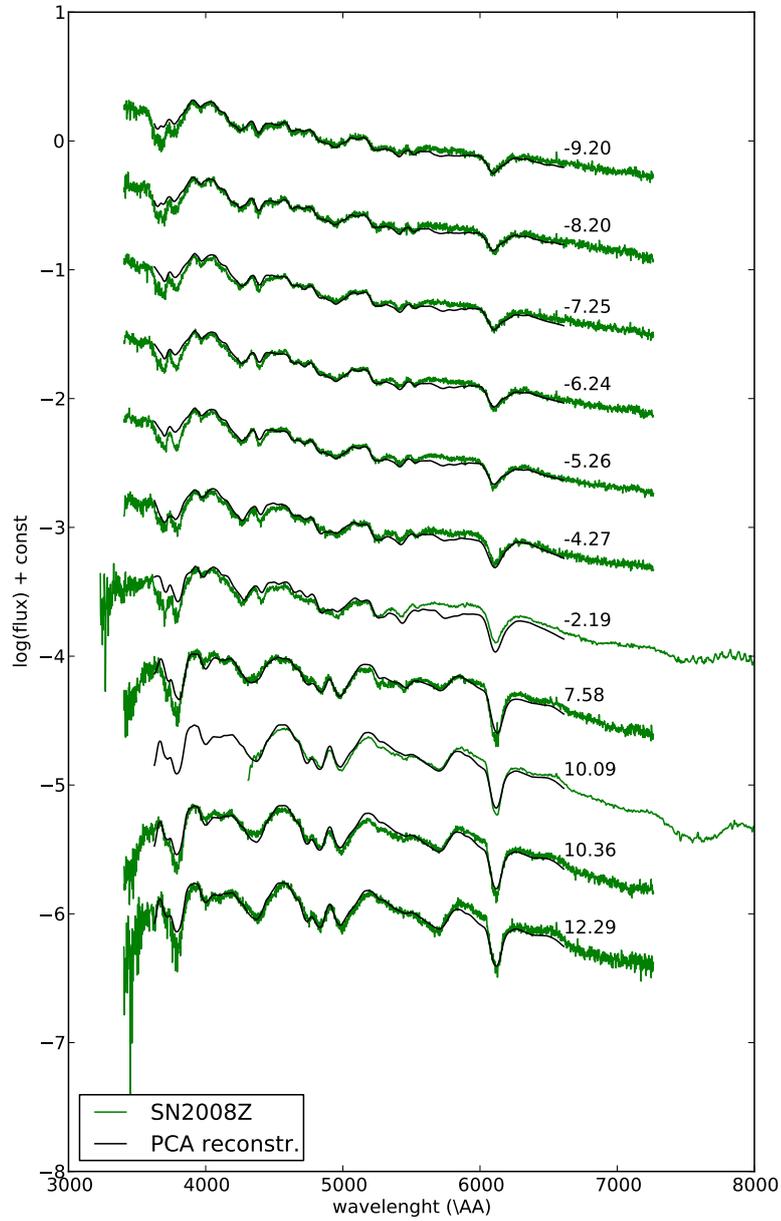


Figure 4.13: As an example, the spectral series of the SN 2008Z is shown in comparison with the spectra reconstructed from the PCA space.

Chapter 5

Comparison with Discrete Observables¹

In the context of the PLS we will now study correlations between the PC space and a few other photometric and spectroscopic quantities. We present a closer look at each of these characteristics and describe in more detail how to obtain such information from the derivative PC space.

The absolute B-band magnitude at maximum is probably the most important quantity for the characterization of SNe Ia. SNe Ia are standardizable candles because a high degree of homogeneity in SN Ia absolute magnitudes can be achieved using simple transformations based on parameters of their light curves. Given the crucial role played by these objects in astronomy and cosmology, a handful of techniques have already been developed aiming at properly standardizing them. The empirical relation between brightness and decline rate demonstrated by (Phillips 1993) is considered to be one of the first standardization techniques for SNe Ia. It is given in terms of $\Delta m_{15}(B)$, which represents the decrease in B-band magnitude at 15 days after maximum brightness. Brighter SNe tend to decline more slowly and consequently present a lower value for $\Delta m_{15}(B)$. This standardization was substantially improved by introducing corrections based on broadband colors (Riess et al., 1996b; Tripp, 1998; Phillips et al., 1999). Ostensibly such color corrections account for extinction from dust, but most likely also contain a hidden color-luminosity correlation intrinsic to the SNe Ia themselves.

For the purpose of comparing models with observations, any successful model should obtain the correct SN Ia absolute magnitudes, and contain the brighter-broader relation. However, in the derivative PCA space the overall flux scaling and broad-wavelength color have been removed, and therefore they are not directly represented in the derivative PCA space. Fortunately there are a number of spectroscopic indicators known to correlate with overall lightcurve peak brightness, width, and color. For instance, Nugent et al. (1995)

¹Part of this chapter is published in Sasdelli et al. (2015)

found that the ratio between the depths of the Si II 5972 Å and the Si II 6355 Å lines correlates with peak B-band absolute magnitude. The pseudo Equivalent Width (pEW) at B-maximum of the Si II 4000 Å line correlates very well with lightcurve width (Arsenijevic et al., 2008; Bronder et al., 2008; Chotard et al., 2011a), as does that of Si II 5972 Å (Hachinger et al., 2006). There is also evidence that the velocity of the Si II 6355 Å line is correlated with the intrinsic SN color (Foley & Kasen, 2011). Since information related to pseudo equivalent widths and velocities will exist, and possibly be enhanced by taking the flux derivative with respect to wavelength, it is quite likely that the derivative PCA space will retain the ability to differentiate between supernovae and models, having different luminosities, light curve widths, and intrinsic colors. Here we apply PLS to explore the presence of such correlations in our derivative PC space.

5.1 Measurement of Discrete Observables

We wish to measure the B-band magnitude at maximum and $\Delta m_{15}(B)$ with the fewest possible modeling assumptions. Therefore, we simply fit a third order polynomial to the B magnitudes measured between -10 and $+25$ days from maximum using errors coming from the noise of the spectra. The fit is evaluated at maximum and at $+15$ days after maximum to obtain the peak B-band magnitude and $\Delta m_{15}(B)$, respectively. Uncertainties come from an error propagation of parameters from the polynomial fit. The V-magnitude at the epoch of B-band maximum is recovered from an analogous fit run on the V-magnitudes. The difference of the two magnitudes at B-maximum gives us the $B - V$ color. The input magnitudes are synthesized from our spectrophotometric time series, using the B and V filter responses given by Bessell (1995). Absolute magnitudes considered here are obtained from the observed apparent magnitudes at B-band maximum assuming Hubble-flow distances, without any extinction corrections. The errors on the absolute magnitudes are computed from the uncertainties in the light-curve fits and added in quadrature to uncertainties due to peculiar velocity of the host galaxies of $\sim 500 \text{ km s}^{-1}$ (Hawkins et al., 2003).

As a point of comparison, we also performed light curve fits using the well-known Spectral Adaptive Lightcurve Template, (SALT2; Guy et al. 2007) code. SALT2 employs an internal model constructed using a linear PCA approach. The model is described by stretch (x_1) and color (c) parameters. The x_1 parameter is analogous to $\Delta m_{15}(B)$, while c is analogous to $B - V$. Here the fits use magnitudes synthesized in the BVR top-hat filters described in Pereira et al. (2013).

Here we focus on three key spectroscopic features: Si II 6355 Å, Si II 5972 Å, and S II 5640 Å. Technical details of the algorithm used to measure their spectroscopic pseudo equivalent widths and velocities directly from the SNfactory spectra are presented in Appendix A.2. Since we do not possess a spectrum at maximum for all of our SNfactory SNe, we determined velocities and pEWs for every available spectra within -7 days and $+7$ days for each SN. These values were then used to perform a linear fit from which we

	PC1	PC2	PC3	PC4	PC5
D(Si II 6355–vel)	0.74	0.58	0.13	0.24	0.21
D(S II 5640–vel)	0.81	0.35	0.35	−0.21	0.24
D(Si II 5972–pEW)	−0.58	0.39	0.21	0.33	0.60
D(Si II 6355–pEW)	−0.12	0.64	−0.38	0.59	0.30
D(<i>B</i> mag)	0.40	−0.63	0.49	−0.32	−0.32
D(<i>V</i> mag)	0.49	−0.60	0.49	−0.21	−0.34
D(<i>B</i> − <i>V</i>)	0.76	0.09	0.43	0.45	−0.13
D(<i>c</i>)	0.76	−0.06	−0.24	0.58	0.11
D(Δm_{15})	−0.53	0.25	0.07	0.39	0.71
D(x_1)	0.66	−0.45	−0.05	−0.26	−0.54

Table 5.1: Directions in PC space found by PLS. Each direction is defined as a linear combination of the first 5 PCs whose coefficients are shown above (e.g., D(Si II 6355–vel) = 0.74×PC1 + 0.58×PC2 + 0.13×PC3 + 0.24×PC4 + 0.21×PC5).

derived the values at maximum and corresponding uncertainties. We required a minimum of three successful measurements in this time window for the SN to be considered for the fit. This method proved to be quite robust. However, it is not capable of distinguishing the HVFs from the normal photospheric component when both are present. Thus, every time we mention independently measured spectroscopic features, we are referring to the velocity of a given line and not its HVFs counterparts.

5.2 Results from Partial Least Square

In section 4.2, we saw that five components are sufficient to address most of the variance in the spectral features present in SNfactory data. Therefore, from now on we will work in a 5D PC space and use PLS to establish correlations between these PCs and other independently measured parameters. Our goal is to demonstrate the potential encompassed by our derivative PC space, which summarizes the evolution of spectral features of a large SN Ia sample. Using the nomenclature of in section 2.5, the PLS technique was used to find the direction in 5D PC space (\mathcal{X}) which best describes each one of the SNe features cited in the previous section (1D - \mathcal{Y}).

Figure 5.1 shows PLS results for the spectroscopic and photometric features discussed in section 5.1, projected – for pedagogical reasons – onto the first 2 PCs. Each one of these lines is obtained from a linear combination of the first 5 PCs, whose coefficients are presented in Table 5.1. In this plot we see the first evidence of important physical information present in the derivative PC space: the connection between the pEW of Si II 5972 Å and $\Delta m_{15}(B)$. As expected from the studies of Nugent et al. (1995) and Hachinger et al. (2006), the direction found by PLS for the pEW of this line is similar to

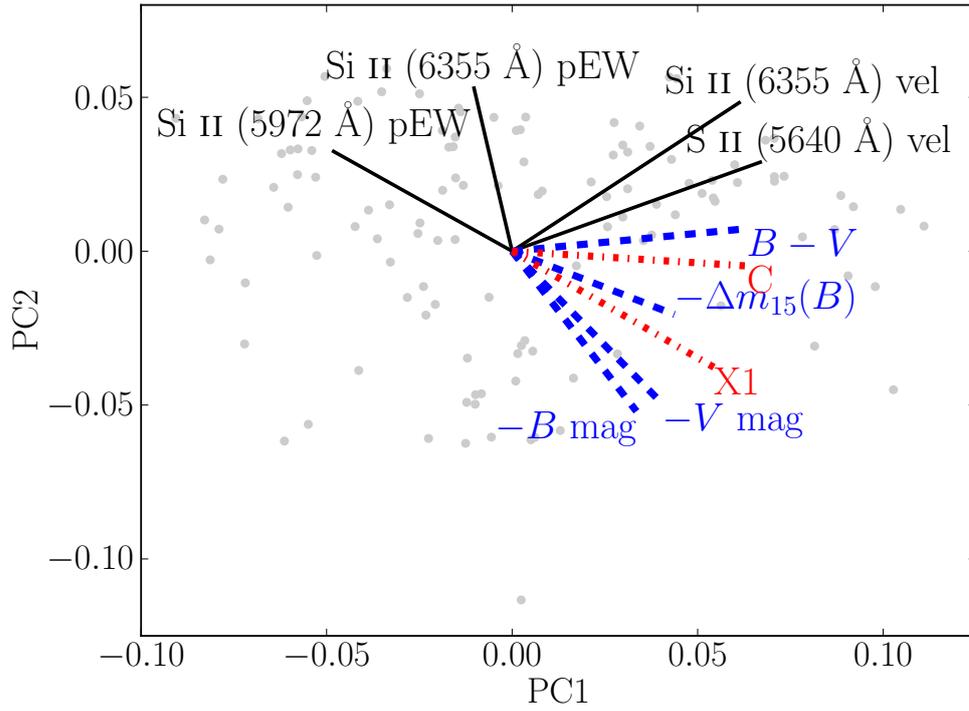


Figure 5.1: The directions maximizing the covariance with various SN parameters derived in a 5-dimensional space and projected into the plane formed by the first two principal components. Gray points are the same as those shown in Figure 4.5. Directions correlated with spectroscopic quantities are coloured in black (solid), photometric quantities in blue (dashed), and results from the SALT2 fit in red (dash-dotted).

	Pearson coeff.	σ_{res}
Si II 6355–vel	0.85	612 km s ⁻¹
S II 5640–vel	0.93	351 km s ⁻¹
Si II 5972–pEW	0.85	4.9 Å
Si II 6355–pEW	0.92	9.9 Å
Δm_{15}	0.78	0.13
x_1	0.74	0.60

Table 5.2: Pearson correlation coefficient for the linear fit between the directions found by PLS and independently measured observables. σ_{res} corresponds to the mean residual between the measured observables values and those determined through PLS. The Pearson correlation coefficient is a measure of the strenght of a correlation and can vary between -1 , 0 , and $+1$ (respectively, perfect anti-correlation, no correlation, perfect correlation).

that of $\Delta m_{15}(B)$ (i.e. opposite to $-\Delta m_{15}(B)$, Figure 5.1). The velocity of Si II 6355 Å is seen to correlate with color, as expected from the study of Foley & Kasen (2011). Interestingly, we also find a strong correlation of the velocity of S II 5640 Å with color. In terms of our PCs, we find that PC1 correlates best with indicators of color.

In Table 5.2 we present the correlations given by PLS for SNe features with each one of the directions highlighted in Figure 5.1. The fact that many important SN features have strong signatures in our new metric spaces gives us confidence that our framework can help us to place better synthetic spectra in relation to their real data counterparts. Next we examine these trends in more detail.

5.3 Spectroscopic Observables in Derivative Principal Component Space

Figure 5.2 shows the correlation between the velocity of Si II 6355 Å at maximum and the corresponding direction found by PLS in PC space. From Table 5.1, we see that it is highly correlated with PC1 and PC2 but not so much with PC3, PC4 and PC5. This is still another angle on the HVFs discussed before: the velocity of Si II is among the persistent features of SNe Ia, and not correlated with the mechanism that gives rise to the HVFs of Ca lines (section 4.3). The few outliers on the high-velocity side of Figure 5.2 are due to strong HVFs of Si II still present around maximum. Their velocity is not predicted by the combination of components that predicts the photospheric velocity, suggesting also that the HVF of Si II is not correlated with the main physics of the explosion and follows the more diverse behaviour of the outer layers.

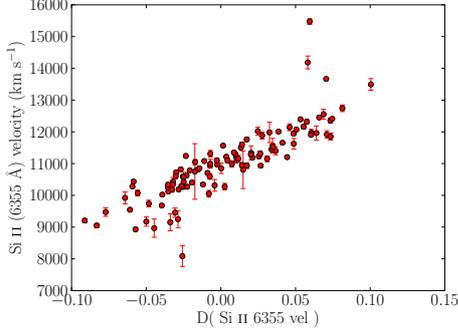


Figure 5.2: Correlation between PLS result and the Si II 6355 Å velocity at B -band maximum. The few outliers on the high-velocity side are due to HVFs of Si (see text).

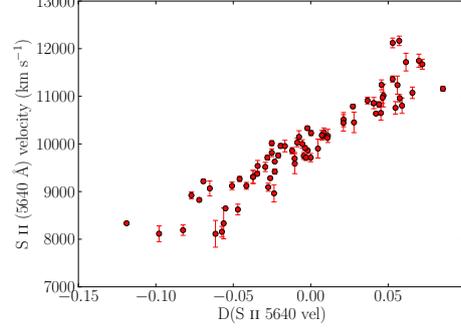


Figure 5.3: Same as Figure 5.2, but for the Si II 5640 Å velocity at B -band maximum.

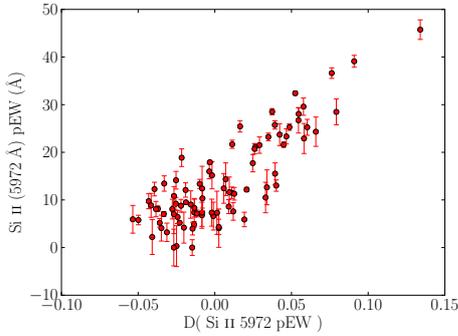


Figure 5.4: Same as Figure 5.2, but for the Si II 5972 Å pEW at B -band maximum.

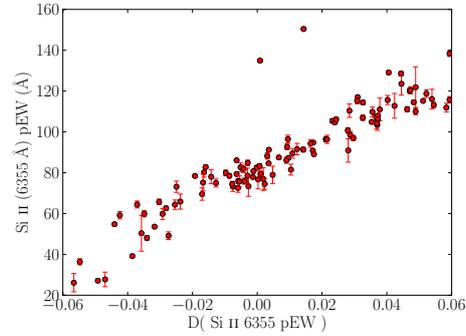


Figure 5.5: Same as Figure 5.2, but for the Si II 6355 Å pEW.

Our ability to describe the velocity of S II 5640 Å using the 5D PC space is illustrated in Figure 5.3. Given the weakness of this line, the quality of the fit is quite impressive (Pearson correlation coefficient (PCC) is 0.93). This is not completely unexpected if one realizes that this line is usually narrower than the saturated Si II 6355 Å line, making a better measure of the velocity possible. More generally, S II lines are not affected by HV features, which can complicate the measurement of the photospheric component. These characteristics suggest that the velocity of S II 5640 Å might present a viable alternative to the Si II 6355 Å line for classification purposes. The S II lines form deep in the ejecta and are good tracers of the photospheric velocity (Blondin et al., 2006). It is expected that for objects with similar luminosities and rise times, a larger photospheric velocity corresponds to a larger radius of the photosphere, a lower radiation temperature and, consequently, a redder color. The ability to extract such an effect from our derivative PC space is very promising as a tool for synthetic spectra characterization. Finally, we emphasize that, although the PC space itself encompasses information regarding the entire time window studied here (-10 to $+10$ days around B-band maximum), the directions obtained by PLS are bounded by the epoch in which the corresponding spectral features were measured. In this context, the correlations presented in Figs. 5.2 to 5.8 are only valid at maximum. An analogous study aiming at a different epoch would require the determination of spectral features at the epoch in question.

Figures 5.4 and 5.5 show the correlation obtained by PLS between the pseudo equivalent widths of Si II 5972 Å and of Si II 6355 Å which are the basis of the Branch et al. (2006) classification scheme. These have Pearson correlation coefficients of 0.85 and 0.92, respectively. This is another indication that information used by others to differentiate between SNe Ia strongly persists in the derivative PCA space.

5.4 Photometric Observables in Derivative Principal Component Space

Having established correlations between spectroscopic luminosity and color indicators and our 5D PC space, we expect to find correlations with B-band peak magnitudes and colors. However, unlike the spectroscopic features discussed above, or the photometric $\Delta m_{15}(B)$ parameter, these are strongly affected by dust extinction and reddening. The information on the amount of the extinction is not present in the dF_{\log} space. This means that observed colors and magnitudes cannot be completely reconstructed using this technique alone. Nonetheless, it is of interest to examine these dust-polluted parameters since their intrinsic behavior is critical for understanding SN Ia physics and standardization for cosmology. This may also allow advances in separating the intrinsic color from dust contributions.

Figures 5.6 and 5.7 show the correlation between the directions in PC space and the observed B and V absolute magnitudes, respectively. All points represent rest frame magnitudes corrected for Milky Way but not for host-galaxy reddening. The well defined

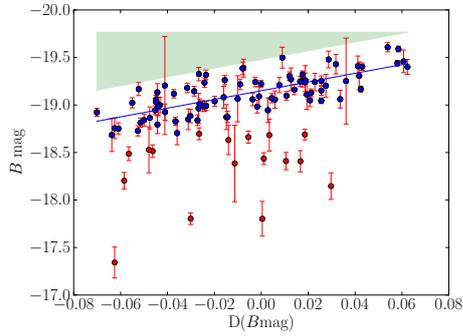


Figure 5.6: Correlation between PLS result and the B -band magnitude. The red points belong to SNe much redder than others with the same spectral characteristics.

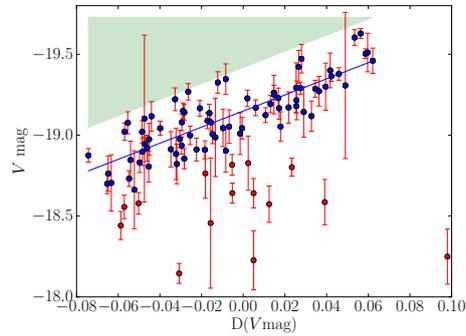


Figure 5.7: Same as Figure 5.6 for V -band magnitude.

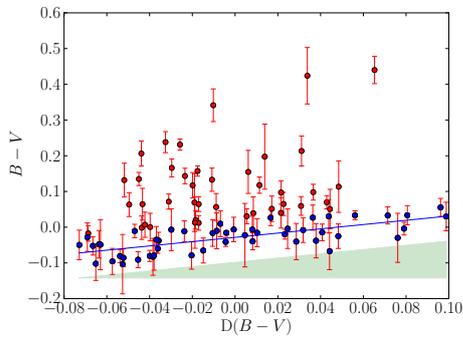


Figure 5.8: Same as Figure 5.6 for the $B-V$ color. The red points belong to significantly reddened supernovae ($E(B-V) \gtrsim 0.1$), and the blue points represent almost unreddened ones.

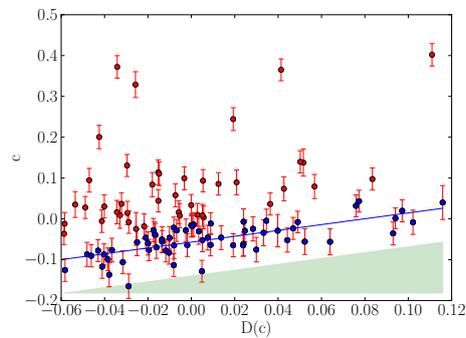


Figure 5.9: Same as Figure 5.8 for SALT2 color parameter c .

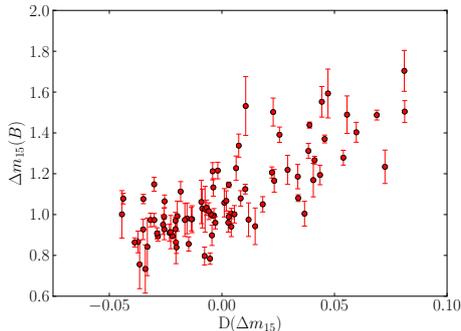


Figure 5.10: Correlation between PLS result and $\Delta m_{15}(B)$. The horizontal axis represents direction in 5D PC space which most correlates with $\Delta m_{15}(B)$ and the vertical axis is the value for this parameter measured from the SNe light-curves.

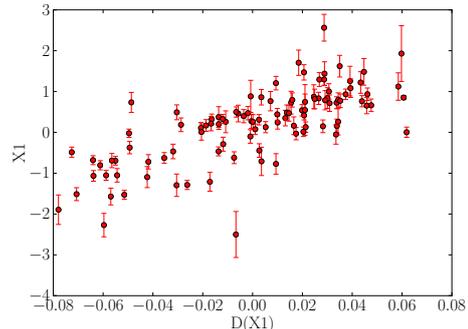


Figure 5.11: Same as Figure 5.10, but for the x_1 parameter of SALT2.

upper envelope situated below the green triangles in each plot suggests a locus potentially dominated by SNe Ia with little extinction. The presence of a slope to this upper envelope versus $D(B)$ and $D(V)$ is likely due to SNe Ia suffering little dust extinction. Because $D(B)$ and $D(V)$ are largely free of the effects of extinction, this strongly suggests that the derivative PC space contains information on the intrinsic luminosity of SNe Ia.

In an effort to find the approximate direction of the luminosity vector, we attempt to isolate the least extinguished SNe Ia under the assumption that brighter SNe Ia have less extinction using an iterative rejection scheme. This type of approach is common when attempting to establish intrinsic peak magnitudes for many SN Ia standardization methods, however, it assumes that $D(B)$ and $D(V)$ impose a sufficient degree of order in the relative SN Ia luminosities, which may be an oversimplification (e.g., Rigault et al. (2013)). (The crispness of the upper envelope is encouraging in this regard.) We applied PLS to the entire data set and then performed a linear fit between the observed magnitudes and the output direction in PC space. Based on this linear fit, only supernovae brighter than the linear fit, or fainter by less than 0.3 mag, are selected for the next iteration. PLS was applied again to the chosen subset and the process was repeated until convergence. The algorithm converged rapidly to a direction that represents the variation of the brightest SNe Ia absolute magnitudes with $D(B)$ or $D(V)$. We found that the output direction in PC space depends only weakly on the criteria used to reject SNe in each iteration. Blue points in Figures 5.6 and 5.7 correspond to SNe selected in the final PLS iteration, the blue line denotes the final linear fit, and red points represent rejected objects.

A similar procedure can be applied to color, using the assumption that the bluest

SNe Ia suffer the least amount of reddening by dust. This assumption is only effective if $D(B-V)$ imposes a sufficient amount of homogeneity in the SNe Ia colors. Again, the crispness of the blue end of the color envelope offers encouragement that this is a sensible approach. Figure 5.8 shows the correlation between $B - V$ color at maximum and the direction in PC space found by the iterative process described above. As in previous plots, each point corresponds to a color measurement without any attempt to correct for host galaxy reddening. The surviving SNe (blue dots) represent objects whose reddening is consistent with the locus of bluest objects to within their measurement errors. Figure 5.9 shows that the SALT2 c parameter has a similar behaviour. This was expected from the existence of a correlation with $B - V$ color at B-band maximum, however the c parameter incorporates the color information at other epochs included in the SALT2 fit. Hence, c corresponds to a more general measurement of the SN Ia color. Here again, because $D(B-V)$ and $D(c)$ are largely free of the effects of extinction, this strongly suggests that the derivative PC space contains information on the intrinsic color of SNe Ia.

Finally, Figure 5.10 illustrates the correlation between $\Delta m_{15}(B)$ and the corresponding PLS result in PC space. The Pearson correlation coefficient between these two quantities is 0.78 (Tab. 2). Discrepancies frequently come from a wrong estimation of the decline rate. Comparing the polynomial fit used to compute the $\Delta m_{15}(B)$ with the SALT2 x_1 , the later usually gives better results. The SALT2 fit takes into account all epochs in B, V and R bands, obtaining a decline rate parameter quite consistent with the one suggested by the EMPCA analysis (Figure 5.11) for most of the objects. This is also reflected in the similar directions found to correlate with $\Delta m_{15}(B)$ and x_1 in Figure 5.1.

We emphasize that the correlations between directions in PC space and global photometric properties like x_1 and $\Delta m_{15}(B)$ represent yet another test of the information encompassed in the metric space. As it was constructed from the entire spectral sequences, it is expected to reproduce such photometric observables even though they were not inserted as features directly into the data matrix. This reinforces our statement that important information is preserved throughout the entire process.

5.5 Infrared Light Curve Properties and Spectral Properties

We collected light curve characteristics published by the CSP collaboration. SN ejecta have a lower opacity in the infrared than in the visible. This permits, with the study of late infrared LCs, to investigate the inner structure of the ejecta. We investigate the correlation between the secondary maximum and the spectral properties of the PCA space with PLS regression.

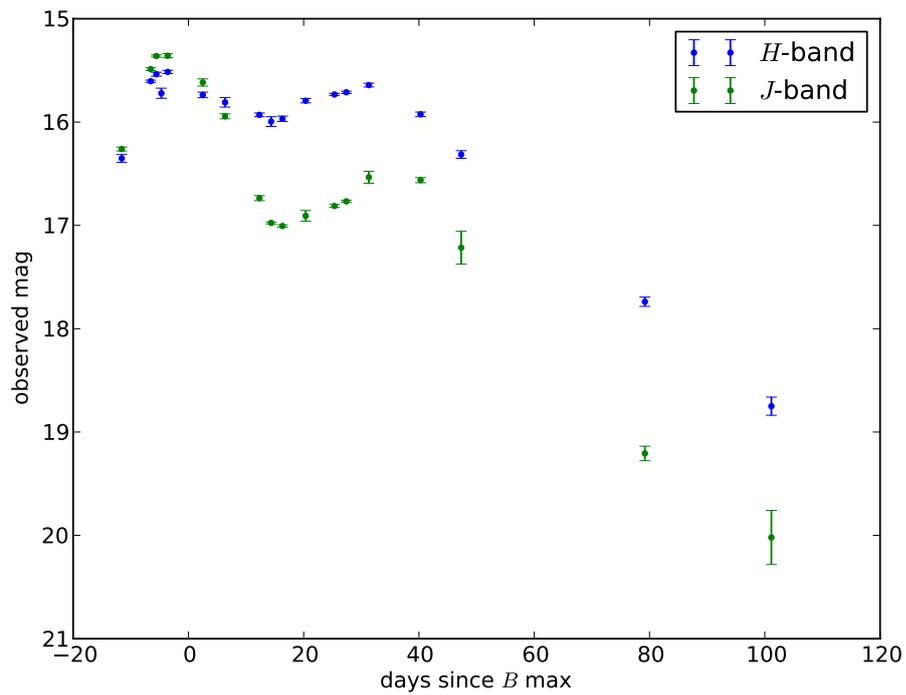


Figure 5.12: The figure shows the behaviour of the NIR light curve of the typical SN Ia SN2007S in the H and in the J bands. The NIR light curves show a second peak at $\sim 30 - 40$ days from the first peak.

5.5.1 Infrared Second Maximum

Infrared light curves of SN Ia have a double peaked structure with a second maximum at around 30 days after the B -band maximum (see the normal SN Ia SN2007S in Fig. 5.12). Kasen (2006) explains the behaviour in the NIR as a combination of the structure of the abundances and behaviour of the recombination. The NIR emission is explained as redistribution of the more abundant blue/UV radiation by fluorescent emission by iron group elements. In a given atomic species, the NIR emission comes from levels with small energy differences between each other. Higher energy levels are generally numerous and closely spaced. At high temperature, the occupation of the levels with large energy increases. Hence, the fluorescence in the NIR of a given species increases with temperature. On the other hand, at even higher temperatures, the species get highly ionized. Hence, the emission in the NIR is strongest close to ionization fronts. The temperature in the ejecta decreases at larger radii and with time. When the ionization front between doubly and singly ionized species of Co and Fe reaches the iron group rich center the emission in the NIR gets enhanced and this forms a second peak in the luminosity.

The luminosity and epoch of the first maximum is known to be fairly uniform among SN Ia and has been proposed to be used as standard candle. The strength and epoch of this secondary maximum is known to be uncorrelated with the intensity of the first maximum of the light curve. Dhawan et al. (2015) show that the epoch and strength of the secondary maximum correlate with $\Delta m_{15}(B)$ and with the pEW of Si II 5972 Å at maximum. This suggests that these parameters are mainly driven by the mass of ^{56}Ni . We explore the correlation of the epoch of the secondary maximum with the PLS regression techniques on the PCA spectral space.

The correlation of the epoch of the secondary maximum with the spectral properties of the PCA space is quite strong (Figure 5.13). Late time infrared is driven by the structure of the inner parts of the ejecta. This means that the variability of the inner parts of the ejecta shows up in the spectra around maximum.

The direction that predicts the time of the second maximum is close to the direction that predicts the pEW of Si II 5972 Å at maximum.

5.5.2 Type Ia Supernovae are “Standardizable” Candles in the Near Infrared

We investigate possible relations between the near infrared (NIR) maximum luminosity and spectral properties. We again use PLS regression between the measured NIR maximum magnitudes measured by Weyant et al. (2014) and our metric space for spectra. The observed H -band magnitudes that we use come from template fitting using Snoopy (Burns et al., 2011). The observed magnitudes have been transformed into absolute magnitudes using the host galaxy redshift and assuming a Hubble constant of $H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$ and $\Omega_m = 0.27$. Ω_m has very little effect at the redshift of our observations ($0.01 < z < 0.09$), the assumption of H_0 adds only a constant to the magnitudes

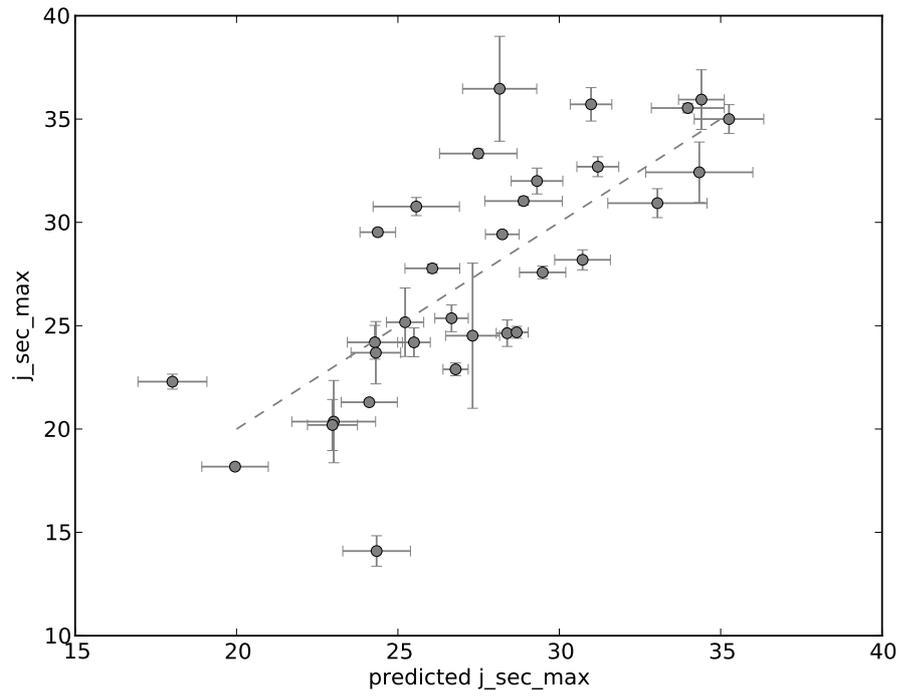


Figure 5.13: The epoch since B -maximum of the secondary J -band maximum compared with the prediction from the PCA space of spectra. The errors on the predictor are computed by k-folding.

without affecting our analysis. The error on the absolute magnitudes is computed adding in quadrature the error on the observed magnitudes and the error due to the intrinsic motion of galaxies. As before, we assume 500km s^{-1} for the peculiar velocity of the galaxies (Hawkins et al., 2003). In the NIR bands reddening is known to give a small contribution and we will not correct for it.

Kattner et al. (2012) showed that the absolute magnitudes in the NIR weakly correlate with $\Delta m_{15}(B)$. We show that there is a relation between spectral properties and NIR magnitudes, in particular on the bright end of the diagram (Figure 5.14). The x-axis shows the predictor of the absolute magnitude in the H -band found by PLS regression. The space of the predictors are the PCA space of spectral series.

The luminosity of SN Ia in the infrared is not uniform, as previously suggested. In particular the high luminosity part of the diagram shows a significant variability that correlates with the spectral properties. This makes it challenging to use the NIR luminosity directly as standard candles. On the other hand, the NIR luminosity can be “standardized” using spectral information, as shown in figure 5.14.

The relation with our predictor space probably has a non-linear nature. The SNe on the bright end of the diagram display a larger spread of luminosities and the dim end shows a smaller variability. The spectroscopic properties that predict the H -band luminosity are shown in Figure 5.15. Spectra are shifted on the y-axis proportionally to the quantity that predicts the H -band luminosity. The spectra on the top are the ones with the brightest NIR luminosity. Brighter SNe with “hotter” spectra are the ones with a larger NIR luminosity. In Figure 5.15 we are restricted to SNe with the spectra at B -band maximum, but our analysis is able to cope with sparse spectroscopy.

The direction that predicts the luminosity in the H -band is close to the direction that predicts the pEW of Si II 5972 Å at maximum. This can be shown by a simple inner product using the metric defined by the covariance of our PCA space. The values of the inner products of the predictor of the H -band and of the predictors of $\Delta m_{15}(B)$ and Si II 5972 Å are, respectively, 0.86 and 0.90. This direction is close to the direction that predicts $\Delta m_{15}(B)$, in agreement with (Kattner et al., 2012). These directions, albeit close, are not equivalent and they are significantly separated from each other. This is unambiguously shown by Fig. 5.16. The distribution of directions that PLS finds for each of these quantities is clearly separated from the distributions inferred for every other. This means that studying our metric space on spectra is not equivalent to study individual spectral indicators. Our method is an alternative to spectral indicators that can systematically study the correlations between spectral properties and light curve properties. From a practical point of view, our metric space takes advantage of spectra in various epochs and not only at maximum. This allows to not reduce the sample when a maximum spectrum is not available.

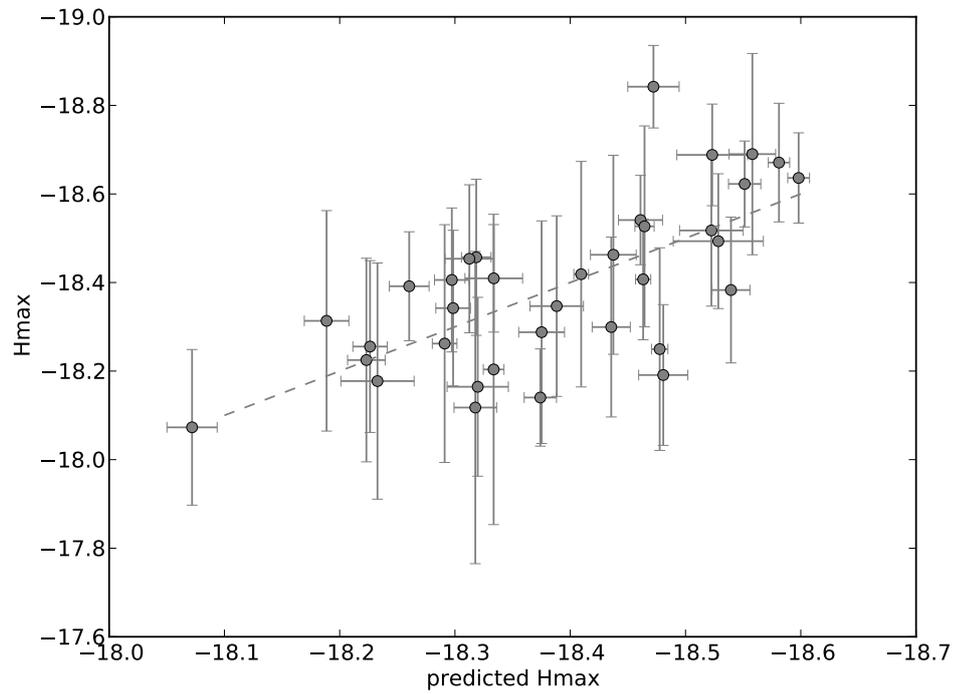


Figure 5.14: The observed absolute H -band maximum plotted against its best predictor from PLS on the PCA space of spectra

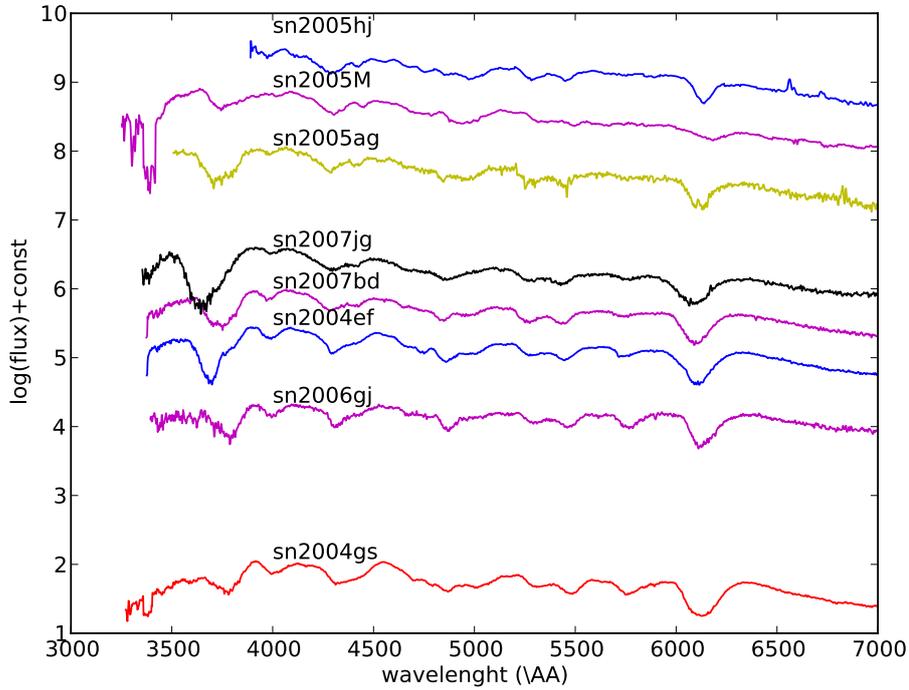


Figure 5.15: SNe with a spectrum at maximum are shown. The spectra are shifted proportionally to the predictor of the H -band magnitude. SNe brighter in the NIR are on top.

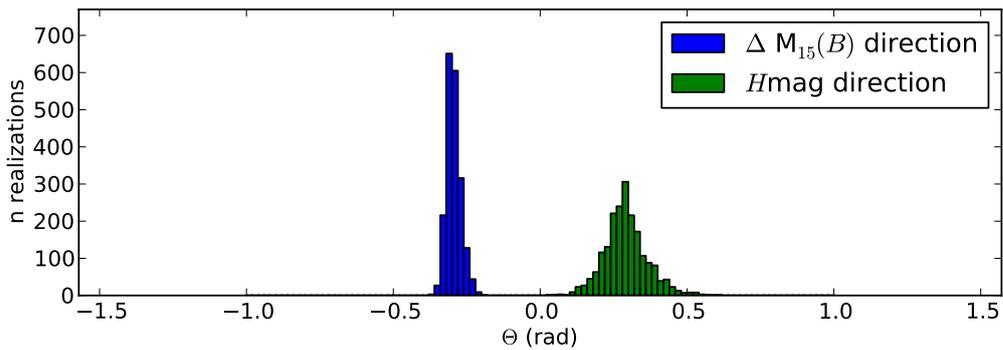


Figure 5.16: The angle Θ between the directions that predict $\Delta m_{15}(B)$ and the H -mag. The two directions are similar, but significantly different. The distributions are sampled with k-folding.

Chapter 6

Comparison with Models

Constructing models from first principles is currently the main strategy for modeling Type Ia supernovae (SN Ia) (Hillebrandt & Niemeyer, 2000).

In doing so, one assumes a progenitor system and explosion scenario and simulates nuclear burning and the explosion in detail. By varying the (physical) input parameters different realizations of every scenario are obtained. Currently investigated models are described in detail in a recent review by Hillebrandt et al. (2013). Since the progenitors of SN Ia are not known, in most cases the various scenarios are simulated for a wide range of reasonable parameter values in order to see if the observed diversity of SN Ia can be explained (Sim et al., 2013; Seitzzahl et al., 2013; Fink et al., 2014). For example, varying the initial mass in case of sub-Chandrasekhar mass double-detonation models changes the predicted luminosity of the explosion (Kromer et al., 2010; Sim et al., 2010; Moll & Woosley, 2013). Alternatively, in some cases specific realizations were studied as possible explanations of unusual events (Pakmor et al., 2010; Kromer et al., 2013).

In a second step, one computes synthetic light curves and time sequences of spectra for the models and compares them with observations. The production of the light from the radioactive decay of ^{56}Ni and ^{56}Co is calculated and the propagation of photons and their escape from the ejecta is computed, in 3-dimensions usually by means of Monte-Carlo methods (Kasen et al., 2006; Kromer & Sim, 2009). However, this approach is computationally expensive and so far only a small part of the parameter space of the presently favored explosion scenarios was investigated. Moreover, it allows little freedom to adjust the resulting synthetic observables to fit the observations. Finally, it is not easy to use observations to guide the explosion modeling. In particular, it is difficult to compare spectra from individual models with individual supernovae on a systematic and quantitative level.

The current approach to test models against observations is mostly done by comparing the light curves of groups of models with the known global properties of light curves of SN Ia. One of the most important and best known of these properties is the Phillips relation, that is, the correlation between the decline rate of the light curve and the luminosity

at peak (Phillips, 1993). Due to the relative simplicity of light curves of SNe Ia it is easy to assess whether realizations of a given explosion scenario follow the Phillips relation or not. However, comparing the global properties of spectra with model predictions is a much harder challenge. Up to now it is done mostly qualitatively (χ^2 by eye) on a case-by-case basis, that is, by comparing a specific model from a given scenario with an individual supernova (or a representative example of a particular class) (e.g. Röpke et al., 2012). It is obvious that within this approach models cannot be tested against empirical relations between different spectral properties and between spectra and light curves of real supernova, which would be more constraining for the models.

Here we are using a different approach. In a first step we construct a 'metric space' for SN Ia spectral time series by means of a *Principal Component Analysis (PCA)* based on a large sample of observed SN Ia (see chapter 2 and section 3.2, respectively, for a description of the method and of the database). With PCA is possible to discover correlations between spectral properties (if they exist) and empirical relations between spectra and photometry can be studied systematically with *Partial Least Square* regression (See chapter 5). Next, the projections of synthetic spectra of models can be computed in the PCA space of the data. It will be shown that this approach allows us to derive constraints for the models in a more systematic way than was previously possible.

6.1 Explosion Scenarios

It is widely accepted that SNe Ia are the result of the thermonuclear explosion of a carbon-oxygen white dwarf triggered by the interaction with a companion star. But beyond this very little is known with certainty. In this section we briefly review the presently favored scenarios (see also Hillebrandt et al. (2013) for a recent review).

Delayed Detonation

This explosion mechanism is usually proposed for single-degenerate systems where the white dwarf explodes close to the Chandrasekhar mass ($1.4M_{\odot}$) after accreting mass from a non-degenerate companion, presumably through Roche-lobe overflow. The matter steadily burns to carbon and oxygen on the surface of the white dwarf increasing its mass until the density at the center is sufficient for the ignition of nuclear burning and a combustion wave to form.

In this class of models it is assumed that in the beginning burning proceeds with a flame speed lower than the speed of sound (deflagration) and incinerates the interior of the star. This phase allows for the white dwarf to expand and decrease the density of the unburned material, a necessary ingredient for the synthesis of intermediate mass elements (IME). It is further assumed that in a next step a transition from a deflagration to a detonation takes place somewhere in the star with a burning velocity now larger than the speed of sound. Whether or not this happens in reality is heavily disputed (Woosley, 2007;

Röpke, 2007; Aspden et al., 2010), but it cures several of the problems of pure-deflagration models, i.e., these models can be brighter and have less unburned carbon and oxygen at low velocity. In fact, after the transition the detonation front quickly burns most the remaining fuel, partially to ^{56}Ni and partially to IMEs. It is this property of the delayed detonation models that brings Chandrasekhar-mass explosions closer to the observed light curves and spectra of normal SN Ia than pure-deflagration models (Schmidt et al., 2010; Poludnenko et al., 2011; Charignon & Chièze, 2013).

Sub-Chandrasekhar Mass Detonations

Sub-Chandrasekhar mass models, i.e., exploding white dwarfs with a mass lower than M_{Chan} , are detonations ignited near the center of the white dwarf. In contrast to the previous case, for these stars the density at the center is not high enough to self-ignite carbon and oxygen but a trigger is needed. A possible mechanism is the so-called double-detonation. A layer of helium-rich material on the surface of the white dwarf may detonate first, for example after it was accreted from a companion (He-)star or during a merger with the companion. The He-detonation will engulf the white dwarf sending shock waves inward which will converge close to the center. Numerical simulations have shown (Fink et al., 2010; Moll & Woosley, 2013) that in the converging shocks the temperature increases sufficiently to trigger a secondary detonation in the C+O fuel. Burning at the lower density of the sub- M_{Chan} white dwarf (as compared to the Chandrasekhar-mass case) produces naturally a large amount of IME (that are seen in the ejecta), and the mass of the initial white dwarf (setting its density) is an excellent parameter to drive the mass of ^{56}Ni and reproduce the variance in luminosity observed in SNe Ia (Sim et al., 2010).

Double-Degenerate Mergers

A scenario completely different from the previous ones is the merger of two sub- M_{Chan} white dwarfs. The orbit of these objects slowly decays through gravitational waves emission until the two stars may merge. If this happens on a timescale shorter than the Hubble time the binary may be a candidate for a SN Ia. The process of merging may trigger an explosion in one or both of the stars immediately, or the less massive white dwarf may get disrupted over a few orbits and a 'hot spot' on the more massive primary triggers a detonation there as in the double-detonation scenario of the previous subsection (Pakmor et al., 2013). Also, as in the case of sub- M_{Chan} detonations, the mass of ^{56}Ni is largely determined by the mass of the primary white-dwarf, and the mass of the secondary and the viewing angle could be additional parameters to explain the diversity within SNe Ia. Alternatively, if the two white dwarfs have comparable masses both stars will get disrupted by the interaction. The lower central density leads to a lower amount of ^{56}Ni and to a significantly lower luminosity. This was suggested as a scenario for the SN 1991bg-like subluminous supernovae (Pakmor et al., 2010).

Pure Deflagrations

This scenario applies for Chandrasekhar mass explosions of white dwarfs, but in this case it is assumed that the transition to a detonation does not happen. Consequently, for equivalent initial conditions, the burning is less complete than in the corresponding delayed detonation models. The production of ^{56}Ni is more limited and this limits the maximum possible luminosity of the scenario. Hence this mechanism can be a possible explanation for peculiar faint SN Ia only.

In contrast, the W7 model (Nomoto et al., 1984) is a parametrized 1D deflagration model. In this model the burning proceeds faster than what happens in modern 3D deflagrations and this allows for a more complete burning of the fuel. It is generally in good agreement with normal SN Ia.

6.2 Models in the Principal Component Space

At our disposal we have series of 3D numerical simulations for all classes of models discussed in the previous section, to be compared with the observations. At first we will check the general consistency of the spectra with observed SNe Ia. To do so, we will use only the first part of our analysis, that is the PCA space. From the position of the models in the 5D PCA space it is possible to find interesting clues about the spectral behaviour of the models. Figs. 6.1 and 6.2 show where models lie in the projections on the first three PCs, the most important ones to describe the observed spectra.

The model classes characterized by the variation of a single input parameter (the white dwarf's mass in case of sub- M_{Chan} model, the number of ignition spots in case of delayed detonations and of deflagrations, the masses of the two white dwarfs in case of mergers) show up in our PC space as chains of points describing curves in the 5D space. Along these curves, the input parameter varies continuously leading to continuous variations of the spectral properties. Most of these models lie well within the the space of observed SNe Ia and cover a fair fraction of their diversity.

More specifically, the sub- M_{Chan} detonations in the range of masses $0.97 \div 1.15$ draw a curved line running clock wise when projected on the first two dimensions (Fig. 6.1) that connects faint 1991bg-like supernovae with normals. The faint model with $0.88M_{\odot}$ marks the beginning of this line, and it is quite far away from most normal SN Ia. Its next neighbors are faint 1991bg-like and 2002cx-like supernovae. This is not unexpected and is a confirmation of the general behaviour of the sub- M_{Chan} models (Sim et al., 2013).

Delayed detonations models lie in a completely different part of the PC space. In these delayed detonation models the initial condition that is varied is the number of ignition spots (N). This affects the strength of the deflagration phase. In the beginning of the deflagration phase the rate of nuclear burning is proportional to the surface of the burning front. A small number of ignition spots means that burning is less complete when the conditions for the transition to a detonation are met. This in turn implies less pre-

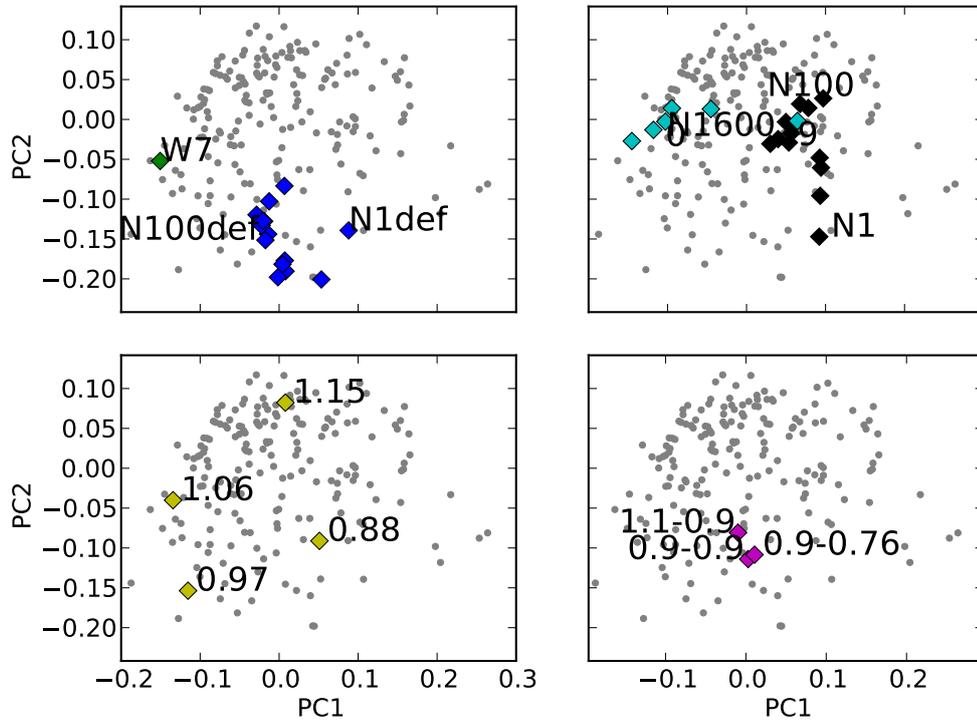


Figure 6.1: The first two principal components of the data with the projections of the analyzed models over-plotted. Most of the models are well inside the PC space determined by the data. Model series are characterized by chains of points in a multidimensional space. The grey dots represent observed SN Ia. The top-left diagram shows the position of W7 on the low luminous edge of normal SN Ia and of the 3D deflagrations together with the faint 02cx-like SNe. The top-right diagram shows the series delayed detonation models with a variable number of ignition spots (black) and models with the composition of N100 and a progressively larger degree of mixing (0 to 9, cyan). The bottom-left diagram shows the sub-Chandra detonations with different initial masses (yellow). The bottom-right panel shows three merger models with different initial masses (magenta).

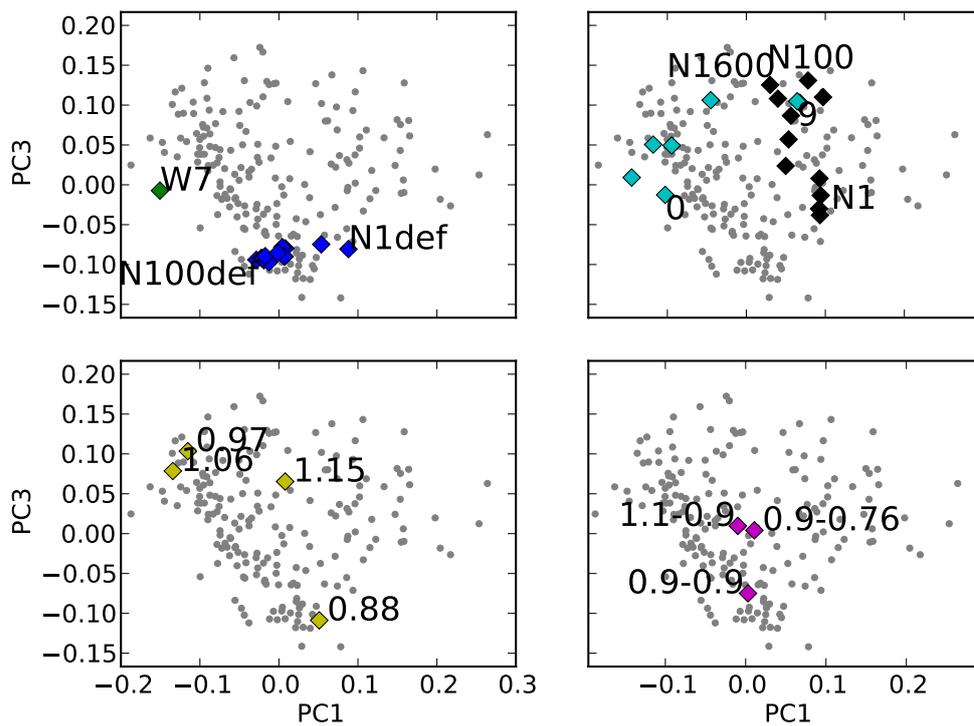


Figure 6.2: The first and the third principal components of the data and of the models over-plotted. The panels are the analogous of Fig. 6.1. deflagrations (top-left), delayed detonation (top-right), sub-Chandra (bottom-left), mergers (bottom-right).

expansion of the white dwarf and a more complete burning in the detonation phase with a larger production of ^{56}Ni . In the first two components the models draw a line that goes from normal SNe with high photospheric line velocities (N1600) down to hot and bright SNe (N1). This behaviour is also in line with what is expected from this class of models.

Klauser (B.A. Thesis, unpublished) created a series of models using N100 (Seitenzahl et al., 2013) as a starting point. First, he constructed a completely stratified model preserving the total masses of the elements present and the density profile. Then, starting from this model he introduced progressively more mixing by convolving the abundances with a Gaussian window. The density profile and the total masses of the different elements were kept constant. This makes the model consistent with the total energy output. The most mixed model has a mixing comparable to N100 but the others have a lower degree of mixing than N100 and they add a component clearly orthogonal to the trend of the delayed detonation models (Figs. 6.1 and 6.2). This finding shows that a mechanism which suppresses mixing in some cases can potentially explain part of the remaining diversity of SNe Ia spectra not matched by varying the number of ignition spots.

Unsurprisingly, the deflagration models stay in the part of the diagram that belongs to the faint 02cx-like SNe (Fig. 6.1). A bit surprising, the prototypical parametrized deflagration W7 model is not close to the bulk of the normal SN Ia but spectroscopically rather agrees with their faint edge.

The two merger models with progenitors of different mass (Pakmor et al., 2012; Kromer et al., 2013) are close to the center of the distribution. The faintest merger of white dwarfs of equal masses (Pakmor et al., 2010) is more separated from the others and closer to 91bg-likes. In fact, this model was designed to fit them.

Many of the scenarios, from a spectroscopic point of view only, could be good candidates for SNe Ia. However, to have a good scenario for SNe Ia, it is important to reproduce not only the right luminosity range, the luminosity-decline rate relation and the proper rise time, but it is also important having consistency between spectral properties and photometric properties. We check this with the aid of PLS regression between the space of spectral properties and a representative set of well studied photometric properties.

6.2.1 $\Delta m_{15}(B)$

The first relation against which we test the models is the correlation between spectral properties and $\Delta m_{15}(B)$. In particular, we study the empirical relation that we found in section 5. This is the analogous to the relation between the ratio of the depth of the Si II 5972 Å and the Si II 6355 Å lines and $\Delta m_{15}(B)$ (Nugent et al., 1995), but studied with the systematic approach of the PLS.

In Fig. 6.3 we show the correlation between the $\Delta m_{15}(B)$ and the direction in the PCA space that predicts it. It is found by PLS and by using public data only. This direction is close to the direction that predicts the equivalent width of Si II 5972 Å (Table 5.1).

The projections of models representative for the different scenarios are over-plotted.

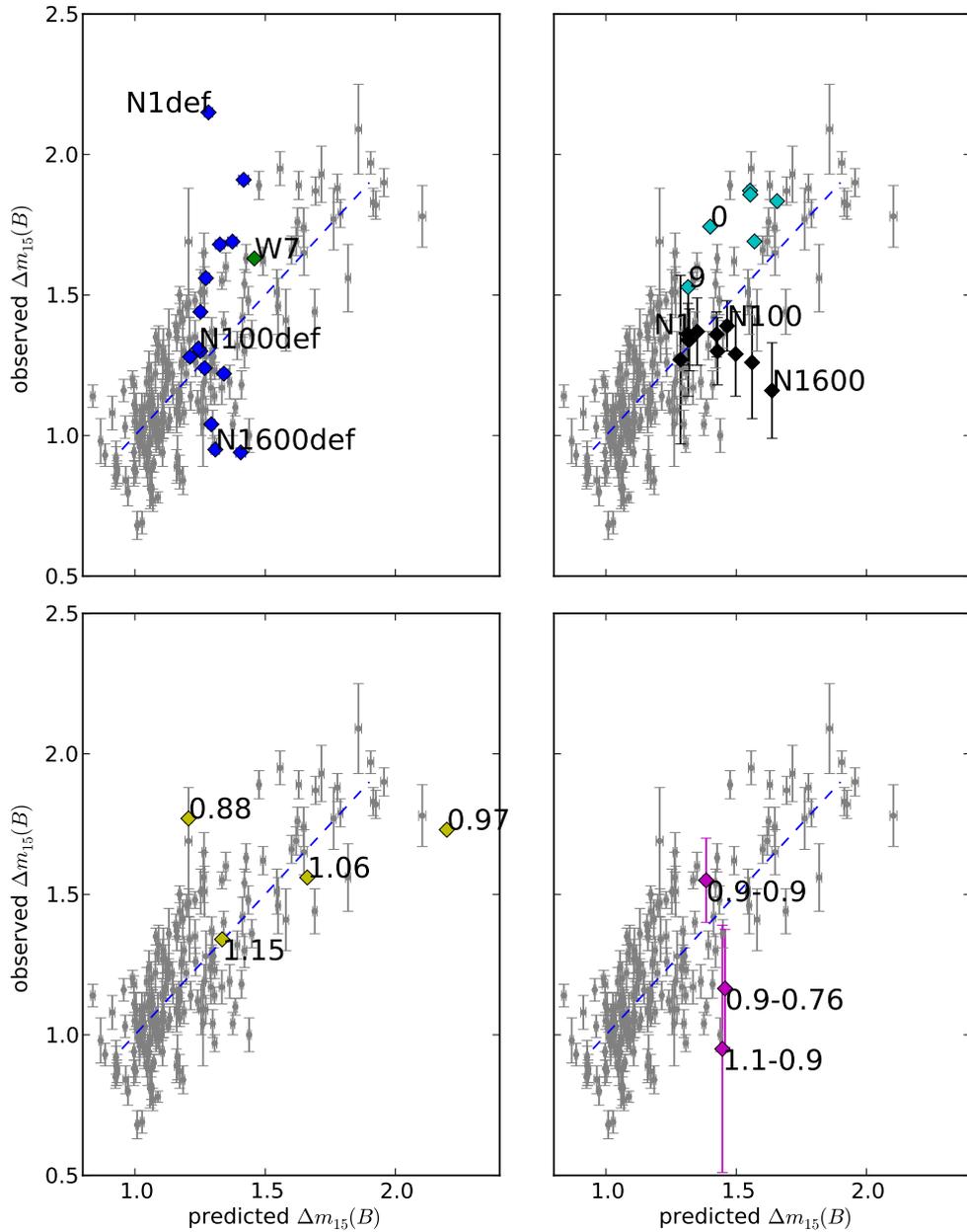


Figure 6.3: The observed $\Delta m_{15}(B)$ relation of observed SN Ia light curves vs. the predictions obtained by using PLS on the PCA space of spectra. The models are deflagrations (top-left), delayed detonation (black) and modified-mixing models (cyan) (top-right), sub-Chandra (bottom-left), mergers (bottom-right). Errors on the predictions come from k-folding (Appendix A.1). The error on the observed $\Delta m_{15}(B)$ of mergers and delayed-detonation models comes from the variability due to line-of-sight effects.

W7, the classical 1D deflagration model from Nomoto et al. (1984), agrees quite well with the relation. In the interval between $0.97M_{\odot}$ and $1.15M_{\odot}$ the sub-Chandrasekhar mass detonation models reproduce the faint part of the relation remarkably well. This is not too surprising since it was shown in Sim et al. (2010) that they follow the Phillips-relation. On the other hand, the relation explored in this section is of a different nature and it is an additional constraint on the models. It is a relation between global spectral properties (predicted $\Delta m_{15}(B)$) and light-curve decline (observed $\Delta m_{15}(B)$). Sub- M_{Chan} detonations in the mass range $0.97 \div 1.15$ follow nicely the observed relation (yellow diamonds in Fig. 6.3).

In principle, the parameter space of the merger model is large but with the three models available to date we can begin to explore it. The brightest merger is clearly below the empirical relation. This means that, for the given spectral properties of the model, its light curve evolution is too slow. In turn, this implies that the opacity of the model is too large which slows down the evolution of the light curve. A likely explanation is that the total mass is too large to reproduce the bulk of normal SNe Ia. This interpretation is confirmed by the qualitatively similar merger model (0.9 and $0.76M_{\odot}$) which matches the relationship much better. The equal-mass merger (0.9 and $0.9M_{\odot}$) lies also on the relation. As discussed before, this relation is very different from the standard Phillips relation since it connects the decline rate with a suitable combination of spectral properties.

Our delayed detonation models cluster in a single area of the empirical relation, but do not show its observed trend. This means that a parameter different from the number of ignition spots is necessary to reproduce the relation. Nevertheless, those models with a number of ignition spots lower than ~ 300 are well within the parameters of observed SN Ia.

Finally, the models plotted as cyan diamonds in Fig. 6.3 have the total abundances of the N100 delayed-detonation model modified by a parametrized degree of mixing (Klauser B.A. Thesis, unpublished). In contrast to the other delayed-detonation models they do follow the empirical relation and models with a low degree of mixing have larger predicted and observed $\Delta m_{15}(B)$. This may indicate that a mechanism that allows for more stratified ejecta may be needed in order to reproduce the observed correlation between $\Delta m_{15}(B)$ and spectral properties with this class of models.

6.2.2 $B - V$ Color

Next we study the consistency of the $B - V$ color of the models at B -band maximum with the observations. The relation between color and spectral properties is similar to the relation between color and velocity of Si II 6355 Å (Foley & Kasen, 2011). Once again, our approach allows for a systematic study of this property and it allows us to use all the information present in the spectra at different epochs, and not only the behaviour of the spectra at B -band maximum.

An important remark here is that in principle our analysis is valid for all SNe Ia. This

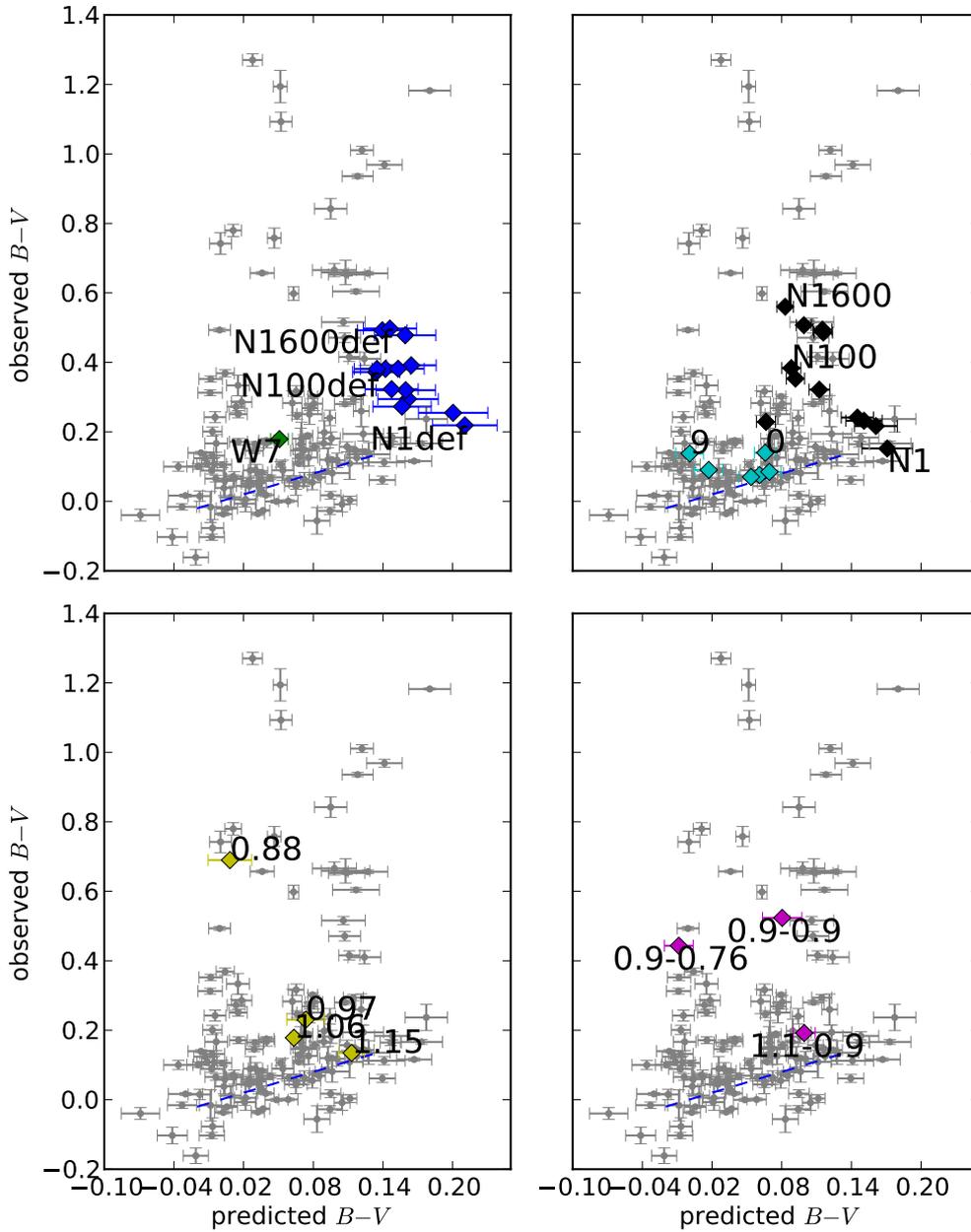


Figure 6.4: The $B - V$ color at B -max (without reddening corrections) and the $B - V$ color predicted from spectral properties using PCA and PLS. The models are deflagrations (top-left), delayed detonation (black) and modified-mixing models (cyan) (top-right), sub-Chandra (bottom-left), mergers (bottom-right). Errors on the predicted colors come from k -folding (Appendix A.1) on the PLS regression analysis.

includes spectroscopically normal ones, 1991T-like, those with broad lines (Branch et al., 2006), and others. But it is not possible to study their color without enough photometric data. In particular the 1991bg-like events present in the sample do not have well enough measured photometry to study their colors with our statistical approach. To quantify which supernovae have a proper prediction of their color typing them is not necessary. It is easy to describe the portion of the PCA space that contains supernovae with no significant reddening by means of the PLS algorithm. This portion of the space includes about 90% of the sample but does not include 1991bg-like and 2002cx-like SNe. With this in mind, the empirical relation between intrinsic color and spectral properties that we find holds for the bulk of SNe Ia, but not for most peculiar ones. Therefore the relation has to hold for models proposed to describe the bulk of the SN Ia but not for rare objects.

Many of the models suffer from being too red compared to the observations (Fig. 6.4). This systematic issue is well known and it may be due to approximations in the radiation transport code. Here, we focus on the trend between intrinsic color and spectral properties that holds for the majority of observed SNe Ia, but does not seem to be clearly reproduced by any of the investigated explosion scenarios for which more than one realization exists.

The delayed detonation models, well known for having high photospheric velocities (Sim et al., 2013), of course place themselves in the right side of the diagram, where the intrinsic color is larger than ~ 0.1 . However, most of these models are still too red. For example, N100, a representative delayed detonation model, is too red by ~ 0.3 mag. It is interesting to note that models with the same composition, but with an enforced stratification, lie very well on the observed relation. The crucial difference in obtaining the right color is most likely the stratification of the inner parts of the ejecta. These modified models have stable iron at the center and ^{56}Ni around it. This changes the way the light is reprocessed and makes the color less red.

The sub- M_{Chan} models show a trend that is opposite to the one observed for normal SNe Ia. Brighter models have IME at higher velocities and they are also bluer. These trends seem to be a robust characteristic of this scenario. More massive progenitors will naturally have higher kinetic energies and ejecta opaque up to higher velocities. At the same time, more massive and brighter models are naturally going to be bluer. Among normal SNe Ia, those with lower photospheric velocities can have both, larger or smaller luminosity.

As previously discussed, there are not enough data available to study the color of 1991bg-like SNe with PLS regression. However, peculiar 1991bg-like supernovae are known to have low photospheric velocities and to be intrinsically redder than normal ones. For them, the sub- M_{Chan} models may be a viable scenario.

The color of the brightest considered merger model is quite right, just a bit too red, and this is likely due to issues in the radiation transport and not in the scenario itself. However, the trend of the merger models with lower masses seems to be orthogonal to the observed relation. The simulation of additional models with intermediate properties is necessary to confirm this trend.

6.2.3 *B* Magnitudes

In order to study the relation between spectral properties and absolute magnitudes we need to have an estimate of the absolute magnitude, independent from the assumption of a Phillips relation. Assuming a Hubble constant of $70 \text{ km s}^{-1} \text{ Mpc}^{-1}$, we use the redshift as a measure of distance. The error of the absolute magnitude is then the error on the observed magnitude with an error due to the peculiar motions of the galaxies added in quadrature. We do not attempt to perform any reddening correction. Similarly as in the case of the colors, we can not study the faintest SNe Ia with statistical methods since we do not have enough 1991bg-like and 2002cx-like supernovae in the smooth Hubble flow.

Sub- M_{Chan} models with initial masses between 0.97 and $1.15 M_{\odot}$ bridge the correct range of luminosities of the bulk of SNe Ia. However, they are orthogonal to the empirical relation between luminosity and spectral properties. The brightest representatives of the SN Ia population are 1991T-like SNe. They are the extreme of the distribution of bright SNe and are characterized by spectra with high temperature and very low photospheric velocities, in contrast to what the sub- M_{Chan} models predict. SNe with a normal luminosity show a diverse range of photospheric velocities that is correlated with color, but not with luminosity as predicted by the sub- M_{Chan} models. To explain the variability shown in the spectra a parameter other than the mass at explosion is needed in this scenario. This parameter, for the same ^{56}Ni mass, needs to increase the total mass of the ejecta to slow down the time evolution. At the same time, it needs to increase stratification, reducing the thickness of the layer rich in IMEs.

Delayed detonation models can also explain the appropriate range of luminosities easily but, like the sub- M_{Chan} models they are mostly orthogonal to the observed relation. The behaviour of models with limited mixing suggests that a mechanism to suppress the mixing of the ejecta can possibly bring the delayed detonation models closer to the observations. Models with high stratification in the ejecta go towards the right side of the diagram. In order to explain the bright SN Ia a high stratification of the ejecta seems to be necessary. This was noticed also by Sasdelli et al. (2014) from the modeling of the luminous SN 1991T. A significant rotation of the progenitor may be a possibility to suppress the mixing and to produce more stratified ejecta.

The W7 model is placed close to the relation. The modern 3D deflagration models, on the other hand, are significantly fainter than the bulk of SN Ia, and can not be used to explain them. Their luminosity, however, is compatible with some of the faint classes such as 02cx-like SNe.

The brightest merger model lies nicely on the relation observed for the bulk of SNe Ia. The lower mass models lie much lower than such a relation, and they cannot be an explanation for it. However, they can explain fainter and rarer objects. To assess if this scenario is a viable explanation for the majority of SN Ia, one has to explore the parameter space close to the bright 1.1 and $0.9 M_{\odot}$ merger to see if the relations hold.

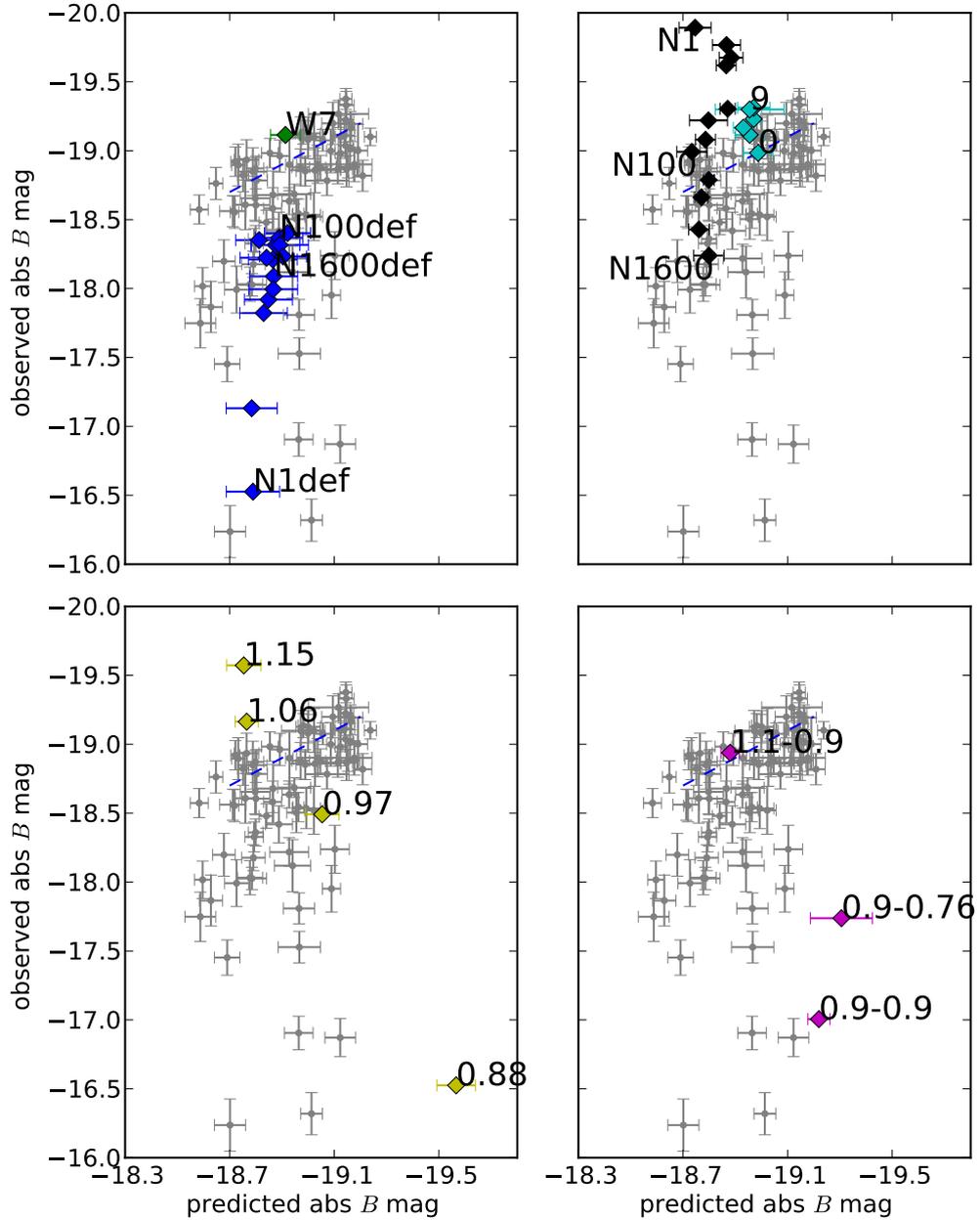


Figure 6.5: The absolute B -band maximum without reddening corrections and the predicted B_{max} using PCA and PLS. The models are deflagrations (top-left), delayed detonation (black) and modified-mixing models (cyan) (top-right), sub-Chandra (bottom-left), mergers (bottom-right). Errors on the predictors come from k -folding on the PLS regression analysis (Appendix A.1). (See Fig. 6.4)

Chapter 7

Multivariate Partial Least Square as a Light Curve Predictor

In this chapter we develop a tool to predict light curves from the information encoded in the spectra. We build upon the work developed in the previous chapters.

From now on we will not use the spectrophotometric time series of SNfactory. The data from SNfactory are very expensive to be obtained. A good tool should be able to exploit the data obtained with the cheaper technique of slit spectroscopy and filter photometry. In this chapter we will show that many of the results obtained by means of SNfactory data can also be obtained from publicly available data described in section 3.2.

We have two types of data and we want to find the latent structures connecting them. The first type of data are time series of spectra. The radiation transport physics that forms the spectra is a complex phenomenon. A lot of information on the physical structure of the ejecta is encoded in the features that show up in the spectra and in the time evolution of these features. On the other hand, information about extinction due to dust along the line of sight is not easy to extract from the spectra. This is due to the difficulty in calibrating data obtained by slit spectroscopy. Also in the case of flux calibrated spectra, it is difficult to separate the variability of intrinsic color and luminosity from the dust-extinction.

The second type of data available is broad-band photometry. The photometry can be used to construct light curves of the SNe in different bands. The data are a more indirect proxy of the spectral energy distribution of the object. Of course some of the intrinsic variability that shows up in the spectra is encoded in the photometry too. After all, photometry measures the flux in a frequency range defined by the filters. But not all the variability of the spectra will clearly show up in the light curves because the effects may get averaged out. On the other hand, extinction information is encoded in the light curves. But also in this case it is difficult to disentangle the intrinsic variability, that has

to correlate with spectral features, and the extrinsic variability due to dust extinction.

We use Partial Least Square regression to find the latent structures connecting these two spaces and to predict the intrinsic component of the light curves from series of spectra.

As shown in chapter 2.2 the data compression of the spectral series is handled with the help of the EMPCA. We need to transform the discrete time series of the photometry to a continuous function. How this is done is the subject of the next section.

7.1 Interpolate the Photometry with Gaussian Processes

We need a robust regression method to fit light curves and color curves. For the Partial Least Square algorithm, we need light curves as a continuous quantity with associated errors. These errors have to take into account the uncertainty of the observed photometry and the sparsity of the data.

SN Ia photometry is typically interpolated with the help of light curve fitters. These algorithms construct a parametrized template for SN Ia light curve. When fitting only B and V bands, they usually employ only two parameters (e.g. SALT2 Guy et al., 2007). The first accounts for the decline rate of the SN (e.g. $\Delta m_{15}(B)$), the second accounts for a color correction using the observed color. This implicitly assumes that SN Ia light curves are described by two parameters only. This is, of course, true as a first approximation, and from only the light curves it is hard to extract many more parameters. But from the study of the spectra it is clear that SN Ia are much more diverse, and we want to take this diversity into account. For these reasons we do not want to use typical SN light curve fitters. With these fitters the result of the fit depends not only on the data of the SN under consideration, but also on the rest of the sample. Of course this makes sense if one wants to calibrate the objects as good as possible, but it does not work if one wants to study the relations between light curves and spectra.

A simple approach could be to obtain the light curves with the least square fit to a polynomial. This introduces biases in the result. Arbitrary choices such as the degree of the polynomial will affect the outcome.

A better approach is the use of a spline fit. This is equivalent to using a series of polynomials smoothly connected together, but also this approach is inadequate for our purposes because it does not give a time dependent uncertainty on the result of the fit.

We are using a powerful regression method based on Gaussian Processes (Rasmussen & Williams, 2005). Gaussian Processes Regression is a more general approach to determine the underlying function from a sparse set of data. The technique assumes that the data are distributed with a Gaussian distribution with infinite dimensionality. In every epoch the outcome of a measurement is assumed to follow a Gaussian distribution. Our set of n photometric measurements can be seen as one realization of an n -variate Gaussian distribution in an n -dimensional space. Now, two epochs close to each other are expected

to be correlated. That is, the luminosity of our SN does not change much in a day. Two observations at different epochs are related by a *covariance function*, $k(t_0, t_1)$, that encodes the relation between the magnitudes at $f(t_0)$ and $f(t_1)$. A safe assumption for the structure of this covariance is:

$$k(t_i, t_j) = \sigma_f^2 \exp \left[-\frac{(t_i - t_j)^2}{2\tau^2} \right] + \sigma_n^2 \delta_{ij} \quad (7.1)$$

The first term means that the correlation is high for observations distant by less than $\sim \tau$ and negligible when the time difference is larger, the second term accounts for the noise. Without noise in the data the correlation between two realizations temporally very close would be equally high to the covariance of an individual realization. Adding noise, the covariance of a given measurement is larger by σ_n^2 . In practice, these coefficients parametrize how slowly the underlying function varies with time (τ), how big the total standard deviation of the magnitudes (σ_f) is, and how much uncertainty is due to the noise of the individual measurements (σ_n). σ_f , σ_n , and τ are called hyperparameters of the model. For a given set of hyperparameters, it is possible to produce a large number of realizations that are likely to reproduce the observations. The average and the standard deviation of these realizations will look like the fits in Figure 7.1. The quality of the fit, however, will be heavily dependent on the choices of the hyperparameters of equation 7.1. For example, if σ_n is set to zero the fit will lie exactly on the photometry overfitting the data. On the other hand, too large a τ will remove small scale variations from the fit, flattening the peaks. The optimal hyperparameters are not chosen by hand but are retrieved by a maximization of the probability $p(\{\sigma_f, \sigma_n, \tau\} | obs)$ of having a certain set of hyperparameters for the given observations.

The result of Gaussian Process Regression on the photometry of the SN 1999dq are shown in Fig. 7.1. Now we have interpolated the luminosities and have sensible error estimates for the fitted light curve. Most importantly, the fits are independent of the behaviour of the other supernovae of the sample.

7.2 Multivariate Partial Least Square

We want to predict the light curves of SNe Ia from time series of spectra. A good tool for this job is Partial Least Square regression. The principles of univariate PLS are explained in 2.5, here we explain the multivariate version of the regression technique.

7.2.1 The Algorithm of Multivariate Partial Least Square

The goal of PLS regression is to predict the \mathcal{Y} space from the \mathcal{X} space to recover their common structure. PLS finds the linear relations that allow to predict a set of quantities in the \mathcal{Y} space, called responses, from the space of predictors \mathcal{X} . The underlying assumption is that every component of the space of responses is a linear combination of the predictors.

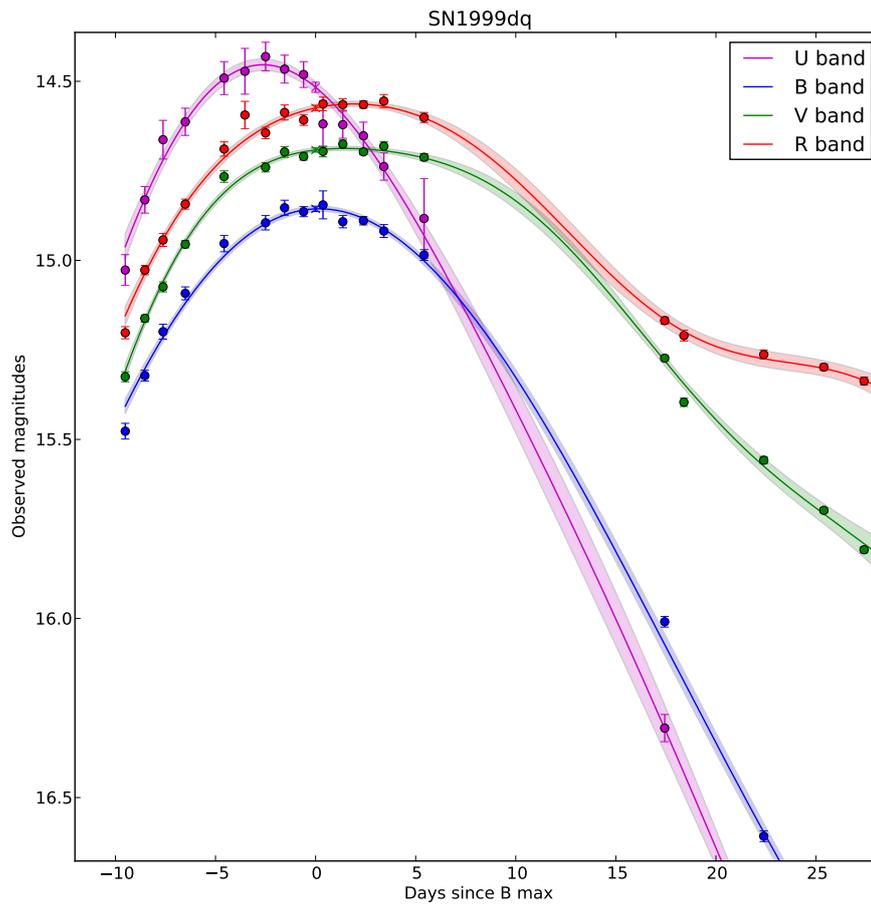


Figure 7.1: Light curve fitting in four bands of the SN 1999dq. The fit is the result of a Gaussian Processes regression.

Mathematically the relation can be written as:

$$\begin{aligned}\mathbf{X} &= \mathbf{TP}^T + \text{residuals} \\ \mathbf{Y} &= \mathbf{UQ}^T + \text{residuals}\end{aligned}\tag{7.2}$$

where \mathbf{X} is the training set of predictors. \mathbf{X} has dimensions $N \times M$ where N is the number of observations (the number of SNe) and M is the dimension of the predictor space \mathcal{X} . In our case M is the dimension of the PCA space built from spectra. The matrix \mathbf{Y} is the matrix of the training set of responses. It has dimensions $N \times L$ where L is the number of epochs sampled from the light curve. \mathbf{T} and \mathbf{U} have dimensions $M \times n$ where n is the dimensionality reduction of PLS, that is the number of components chosen to explain the relation between the space \mathcal{X} and the space \mathcal{Y} .

\mathbf{T} represents the projections on the latent structures defined by the matrix of the weights \mathbf{W} :

$$\mathbf{T} = \mathbf{WX}.\tag{7.3}$$

The scores (\mathbf{T} and \mathbf{U}) have the property that they reproduce well \mathbf{X} and \mathbf{Y} , and the x-scores (\mathbf{T}) are good predictors of \mathbf{Y} :

$$\mathbf{Y} = \mathbf{TQ}^T + \text{residuals},$$

when the residuals of the prediction have to be small.

The relation between the two spaces is best explained by Fig. 7.2, from Wold et al. (2001). The dimensionality of the initial space gets reduced to n , the dimension of the space of the latent variables, by the matrix \mathbf{W} (equation 7.3). In Fig. 7.2 $n = 3$. Then, the matrix \mathbf{U} is responsible to predict the space \mathcal{Y} through equation 7.2. In our case, the “structure descriptors” are the spectral series and the “activity measures” are the light curves. The variables are the coefficients of the PCA space of the spectra, the observations are the different SNe, $\{t_1, t_2, t_3\}$ are the latent variables, \mathbf{Y} is the matrix of the observed light curves, M is the dimension of the range of epochs included in the light curves.

Multivariate PLS is particularly recommended when there is a high correlation between the responses. This is the case for light curves.

The decompositions of \mathbf{X} and \mathbf{Y} are chosen to explain as much as possible of the covariance between the two datasets.

A simple algorithm to compute the weight matrix (\mathbf{W}) and the scores (\mathbf{T} and \mathbf{U}) proceeds as shown in Algorithm 2. This algorithm is implemented in the *scikit-learn* statistical suite (Pedregosa et al., 2011).

7.3 Predicting Light Curves and Color Curves from the Spectra

In this section we use multivariate PLS regression to find correlations between photometry and spectral properties. The spectral properties are described by the coefficients of the

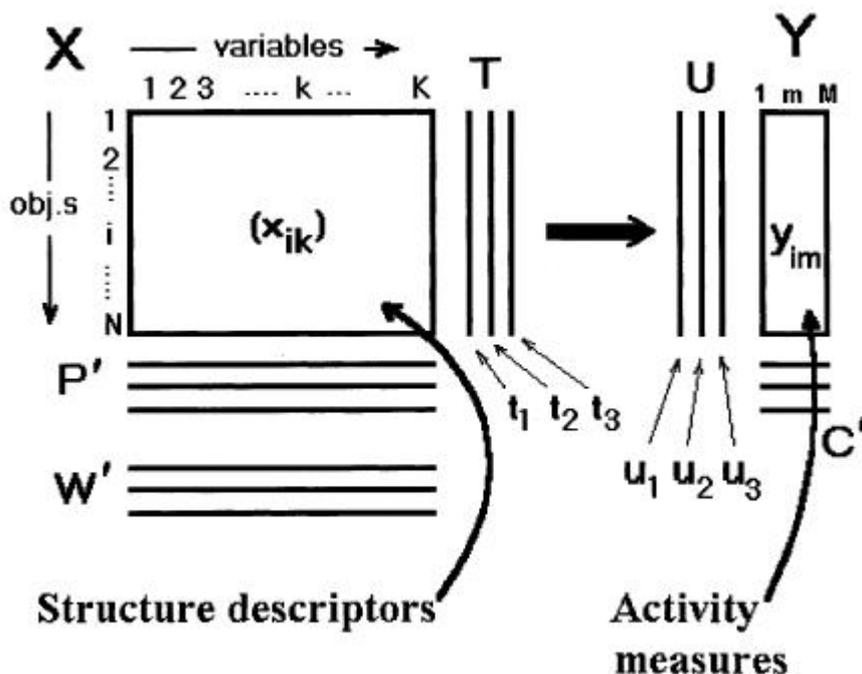


Figure 7.2: The relation between the space of predictors and the space of the responses explained in a graphical form. In our case, the structure descriptors are the spectral series and the activity measures are the light curves. In Wold's example \mathcal{X} and \mathcal{Y} are, respectively, descriptors of chemical structure and measurements of biological activity.

Algorithm 2 Partial Least Square algorithm

1. Assign $\mathbf{X}_0 = \mathbf{X}$ and $\mathbf{Y}_0 = Y$ (first iteration)
 2. repeat n times (the chosen dimensionality reduction)
 - Compute the SVD of the matrix $\mathbf{X}_n^T \mathbf{Y}_n$
 - Compute the first left singular vector (w_n) of the matrix $\mathbf{X}_n^T \mathbf{Y}_n$.
 - Compute the first right singular vectors (v_n) of the matrix $\mathbf{X}_n^T \mathbf{Y}_n$.
 - Compute the n th X-score: $T_n = \mathbf{X}_n w_n$
 - Compute the n th Y-score: $U_n = \mathbf{Y}_n v_n$
 - Deflate the \mathbf{X} matrix: $\mathbf{X}_{n+1} = \mathbf{X}_n - T_n P_n^T$. It is not necessary to deflate \mathbf{Y} .
-

PCA space constructed from spectral series (chapter 2.2). The coefficients from the PCA encode the variance within Type Ia SNe spectral series in a handful of numbers. The good quality of the reconstruction of the spectra prove that this decomposition encompasses the variability of Type Ia spectra. This space, by construction, does not include reddening.

The intrinsic colors and absolute magnitudes are a function of the spectral series. Physically, a given spectral series is expected to have a unique possible intrinsic color and luminosity behaviour. That is, the luminosity and color are expected to be a function of the components of the PCA space constructed from the spectra. We use PLS regression to extract this function. Using this approach, we make the implicit assumption of linearity. That is, the intrinsic colors and absolute magnitudes are assumed to be a linear function of the components of the PCA space of the spectra.

The observed colors are subject to reddening. To find the intrinsic color one has to find the locus of the bluest SN for every point of the PCA space. We use an approach similar to what was used in chapter 2.5 to select SNe with marginal reddening. In that case we had only one parameter for the threshold of maximum reddening to select the SNe, here we have more complicated color curves and light curves. Additionally, we are using public available photometric data that have more diverse errors than SNfactory data. We select supernovae for the PLS regression if they have a reddening lower than a given threshold above the PLS prediction in at least a minimum range of epochs and with errors lower than a given value in this range of epochs. The PLS algorithm and the selection is run a number of times until the solution has converged. The supernovae in each iteration can be selected or deselected. These three parameters are validated through cross-validation.

7.3.1 Predicting the Light Curve from the Spectra

We apply the approach to the B -band light curves between -5 days and $+35$ days from maximum. The aim is to predict the absolute magnitude curve using the spectra of the supernova. This is done free from assumptions on reddening laws and extinction. First of all the observed magnitudes need to be scaled to the reference frame. This is simply achieved using the host-galaxy redshift and assuming a smooth Hubble flow. The observed magnitude needs to be decreased by the distance modulus:

$$M_{abs} = M_{obs} - \mu$$

where, assuming an homologous expansion for the Hubble flow:

$$\mu = 5(\log_{10}(zc/H_0) + 5)$$

The peculiar motion of the galaxies will add an error to this estimate. The error on the absolute magnitudes due to the dispersion of the peculiar velocity $\sigma(v)$ of the galaxies becomes:

$$\sigma_{\mu} = 5\sigma(v)/(\ln(10)zc).$$

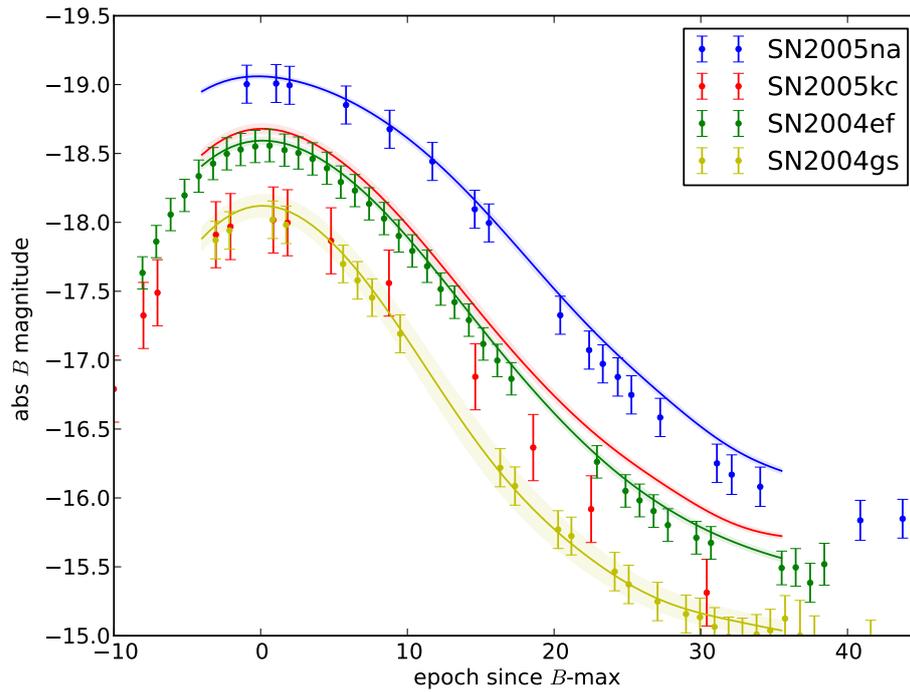


Figure 7.3: The photometry of some SNe with also a good spectroscopic coverage are shown. The curves are the corresponding prediction due to PLS regression.

In Fig. 7.3 we show light curves of a few SNe with excellent photometric coverage. The points are the original photometry, corrected for the Hubble flow, but with no attempts to correct for reddening. The solid curves show the corresponding PLS predictions. The colored area represents the uncertainty of the prediction calculated using k-folding. The photometric data and the predictions match nicely. SN 2005kc, however, has a luminosity significantly lower than is predicted. This can be explained by reddening of this individual object. As was shown in Sec. 4.2, reddening does not influence the components of the PCA space. A supernova with high reddening does not have a PLS prediction different from a supernova with no reddening. From this we deduce that the mismatch between the prediction and the observations comes from dust extinction. The difference between the curve and the data is an estimate independent from assumptions on the nature of the reddening law or amount of reddening. Clearly, this estimate can be calculated at different epochs. Under the assumption that the amount of extinction does not vary with time, we expect that the luminosity deficit of a supernova stays constant. We check it in Fig. 7.4. It shows the extinction in the B band ($E(B)$) at maximum and at +10 days for the supernovae in our sample. The extinctions are consistent between each other, which implies that the extinction does not vary significantly after maximum, and confirms the reliability of the method.

The Phillips-Relationship

Type Ia SNe luminosity is known to anti-correlate with the decline in luminosity after maximum (Phillips, 1993). With the PLS regression we have an estimate for the B -band peak luminosity independent from reddening assumptions. In Fig. 7.5 we show the relation between our estimate and the $\Delta m_{15}(B)$, the difference between the magnitude at maximum and at +15 days. Many of the known characteristics of SN Ia show up in this diagram. On the bottom right are the faint and fast declining 1991bg-like ones. There are only a few objects of that kind in our sample, hence the errors on the predicted magnitudes are large. Spectroscopically normal SNe show a wide range of luminosities and decline rates. On the tip of the relation are the luminous 1991T-like SNe. An interesting “outlier” of the Phillips relationship is the SN 2001by. It is the left-most point in the figure, with a $\Delta m_{15}(B) = 0.7$ (the slowest of the sample). This SN is so extreme that it was clearly recognized as an outlier by Krisciunas et al. (2011) as too faint for its decline rate. Our analysis nicely confirms it and shows that such outliers are not uncommon. With larger $\Delta m_{15}(B)$ their intrinsically lower luminosity is hard to be distinguished from reddening.

7.3.2 Predicting the Color Curve from the Spectra

In this section we apply the PLS regression method on $B - V$ color curves. Differently from magnitudes, the colors are not affected by the distance, and can usually be measured precisely in nearby SNe. On the other hand, the intrinsic variance of colors is smaller than

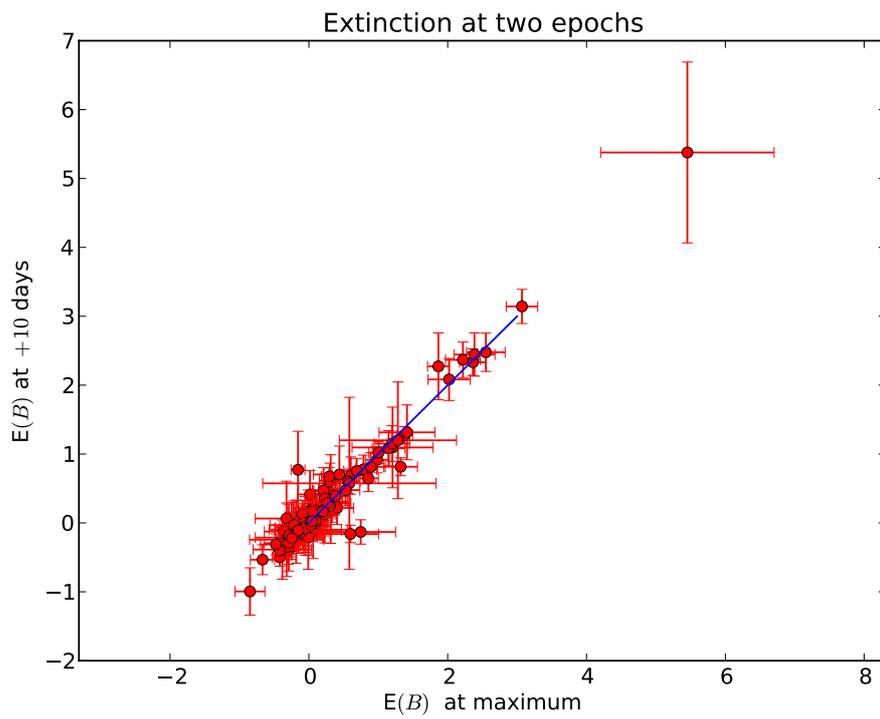


Figure 7.4: The figure shows the $E(B)$ at maximum and at +10 days obtained by PLS regression. A diagonal is overplotted as a reference.

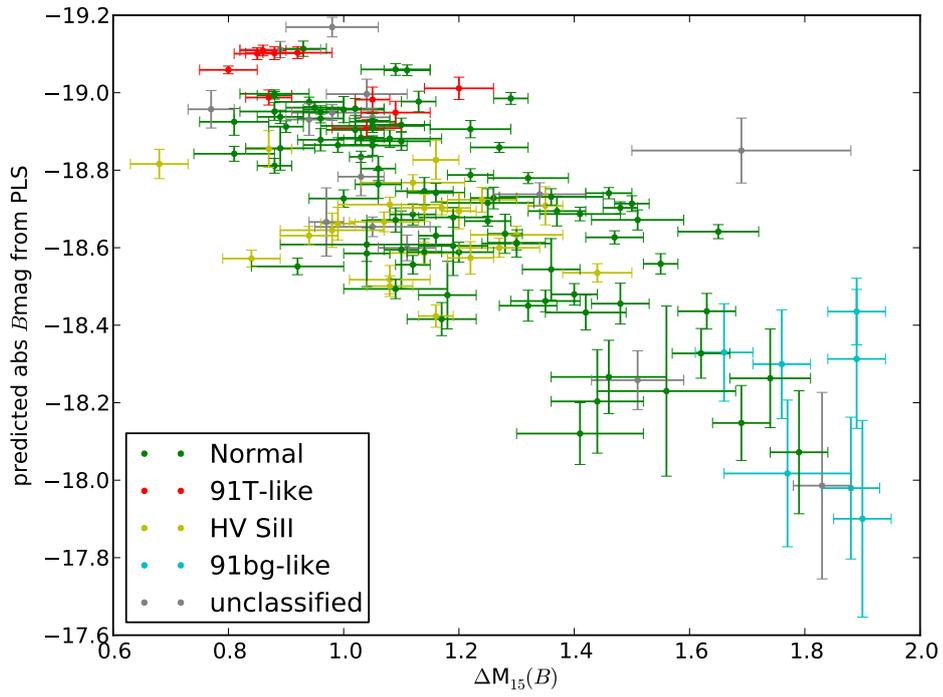


Figure 7.5: The Phillips relationship between the B -band luminosity at maximum predicted by PLS and the decline rate of the light curve. The different subclasses of SN Ia are colored.

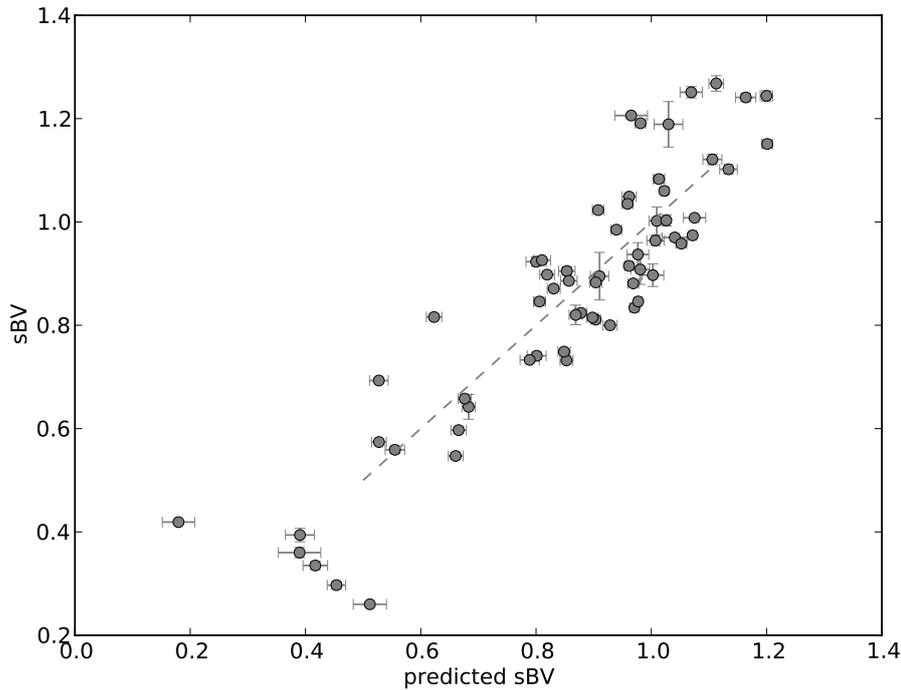


Figure 7.6: The relation between the epoch of the $B - V$ maximum in units of 30 days (s_{BV} , Burns et al., 2014) and its prediction from PLS regression performed on the PCA space of spectral properties.

the variance in the magnitudes. This makes the regression task similarly complicated.

First, we want to show that important characteristics of the $B - V$ color curve are retained in the PCA space of spectra. The time between the B maximum epoch and the maximum in the $B - V$ color curve is a reddening independent quantity that have been suggested for SN Ia classification (Burns et al., 2014). This maximum happens usually at about +30 days after maximum. This time (rescaled in units of 30 days by Burns et al., 2014) is shown to correlate with the spectral properties encoded in our PCA space. Simple univariate PLS regression between the PCA space and this color curve indicator shows an excellent correlation (Fig. 7.6).

In sec. 5.3 we showed that the color at maximum has a correlation with spectral properties. Here we generalized the approach with the help of multivariate PLS. Fig. 7.7 shows some observed color curves (points with errorbars) together the corresponding predictions from the spectra using PLS. It is evident that the colors are quite uniform at maximum, but they have a large spread of properties. SN 2005kc has clearly a significant

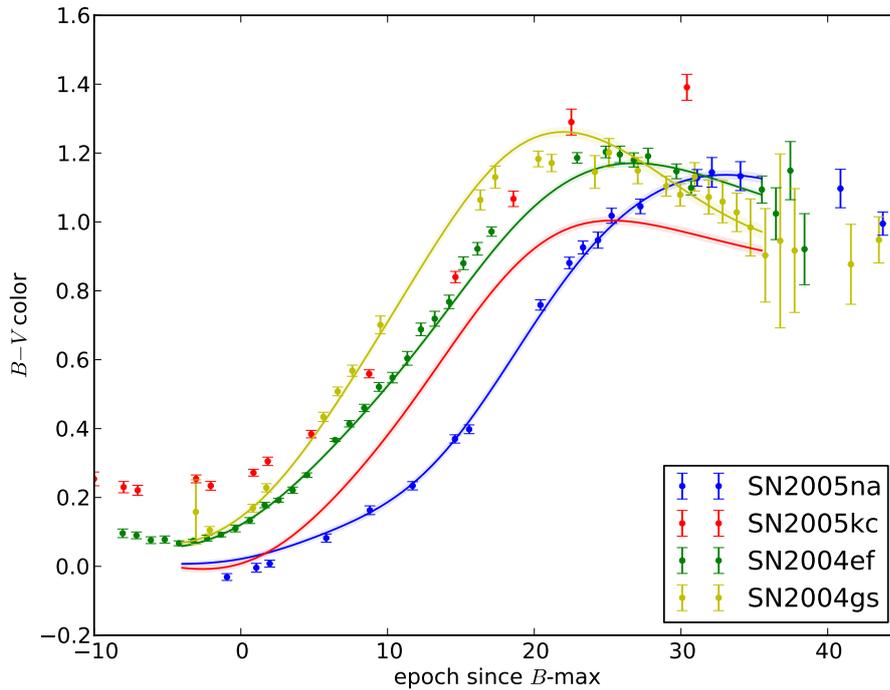


Figure 7.7: The $B - V$ color curves predicted from the spectra of a few SNe with excellent photometric coverage are compared with the color curves predicted from the spectra by means of PLS regression.

color excess in comparison with the prediction from its spectra. This is consistent with a significant amount of reddening.

Similarly to before, we want to check that the PLS regression is catching the intrinsic color variability. Under the assumption that the amount of reddening does not vary with time, we check that the color excess attributed to dust is constant at different epochs. Fig. 7.8 shows the color excess at maximum and at +10 days. The remarkable consistency between the two quantities supports that PLS regression properly predicts the large majority of intrinsic color variation and that the amount of reddening does not vary in this range of epochs.

7.4 Studying the Extinction-Reddening Relation

The extinction due to dust affects more shorter wavelengths. This means that the increase in the magnitude in the B band ($E(B)$) will be larger than the increase in the V band.

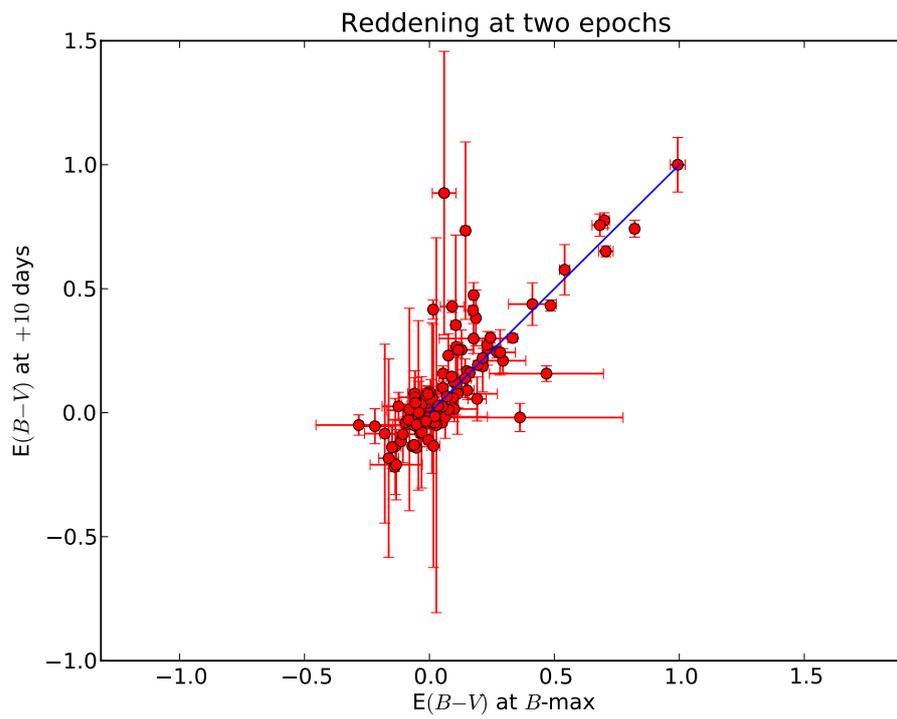


Figure 7.8: The figure shows the $E(B - V)$ predicted by multivariate PLS regression at maximum and at +10 days. A diagonal is overplotted as a reference.

Being a relative measurement, observing colors and measuring the color excess ($E(B-V)$) is typically much easier than measuring the absolute magnitudes. Hence, usually the relation between the extinction in these bands is parametrized as:

$$E(V) = R_V E(B - V).$$

Typical values for R_V measured in our galaxy vary between 2.1 and 5.8 with 3.1 being the most common value (Cardelli et al., 1989; Draine, 2003).

Fig. 7.9 shows the color and magnitudes excess from the PLS reconstruction. The relation is remarkably tight, which means that the value of R_V is remarkably constant for all SN Ia. This is an excellent news for the use of these object as distance indicators. To calibrate SN Ia it is necessary to measure how much extinction is present in front of the SN. And the only practical way of measuring the amount of dust at cosmological distances is through the color excess. Having a constant R_V is always implicitly assumed by light curve fitters. Moreover, frequently SN Ia are associated with a low R_V (frequently < 2) that does not show up in galactic reddening (e.g. Conley et al., 2007; Mandel et al., 2011). This have been explained by peculiar environments around SN Ia (e.g. Goobar, 2008) or, alternatively, as an indication of a peculiarity of the Milky Way dust. On the other hand, techniques based on spectral features return values of R_V much more similar to what is typical in the Milky Way (Chotard et al., 2011b). The result of an orthogonal distance regression fit to the relation of Fig. 7.9 returns an $R_V = 3.06 \pm 0.24$. This is perfectly consistent with the prototypical extinction law (3.1) found in our galaxy.

Fig. 7.10 shows a zoom on the SNe with small reddening. The tightness of this part of the correlation supports that the majority of the intrinsic color and luminosity variability have been taken into account properly by the PLS method.

It is impossible to discriminate between spectroscopically different SNe by using only light curves. However, two SNe with different spectral characteristics may have different intrinsic luminosities and intrinsic colors. The similarities of the light curve and color curve shapes may suggest that the differences are due to dust extinction. The ratio between this “missing luminosity” and “color excess” happens to be significantly lower than what is due to typical dust. This lead to low estimates for the R_V of SNe Ia.

From a physical point of view, a normal R_V for SNe Ia means that the dust in front of the majority of these object is quite normal. This suggests that the majority of their extinction is caused by simple interstellar dust in the host galaxy and it is not related to the progenitor of the SN.

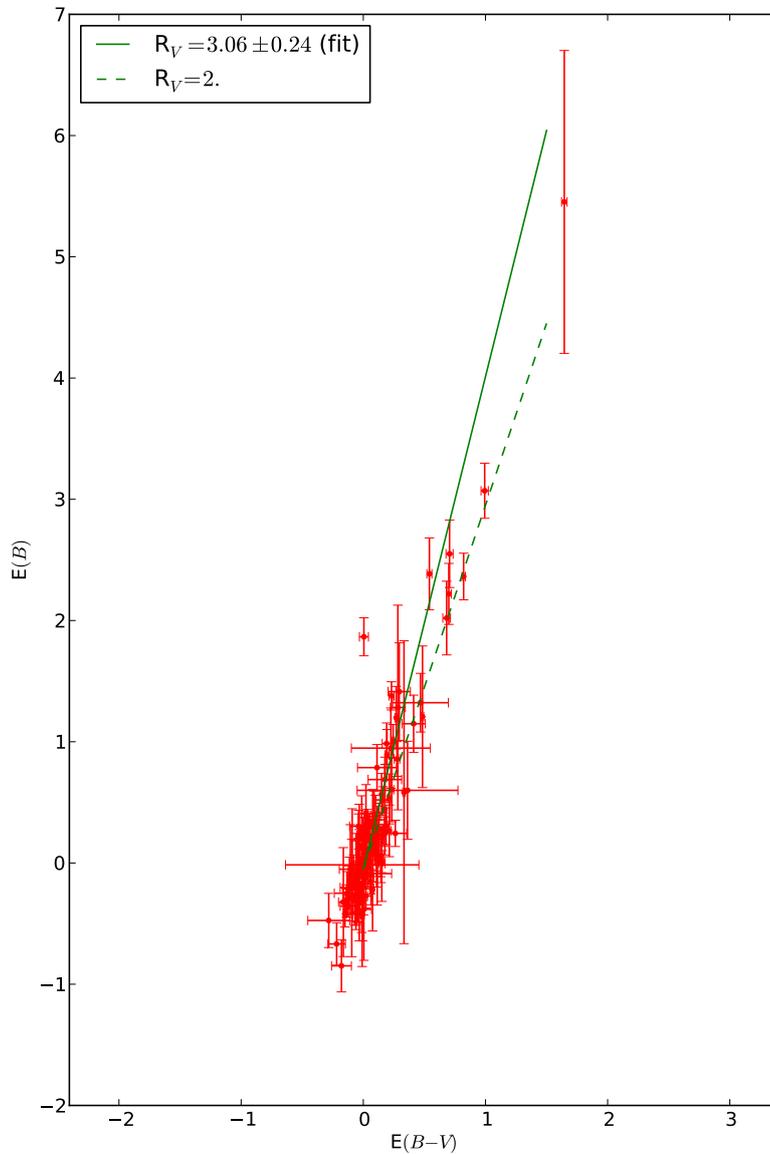


Figure 7.9: The plot shows the relation between the extinction in B and the reddening in $B - V$ at maximum. The solid line shows the result of a fit of the relation. The dashed line, as a reference, is the result of a low R_V . The errors come from the measurement errors and from statistical errors (k-folding) added in quadrature.

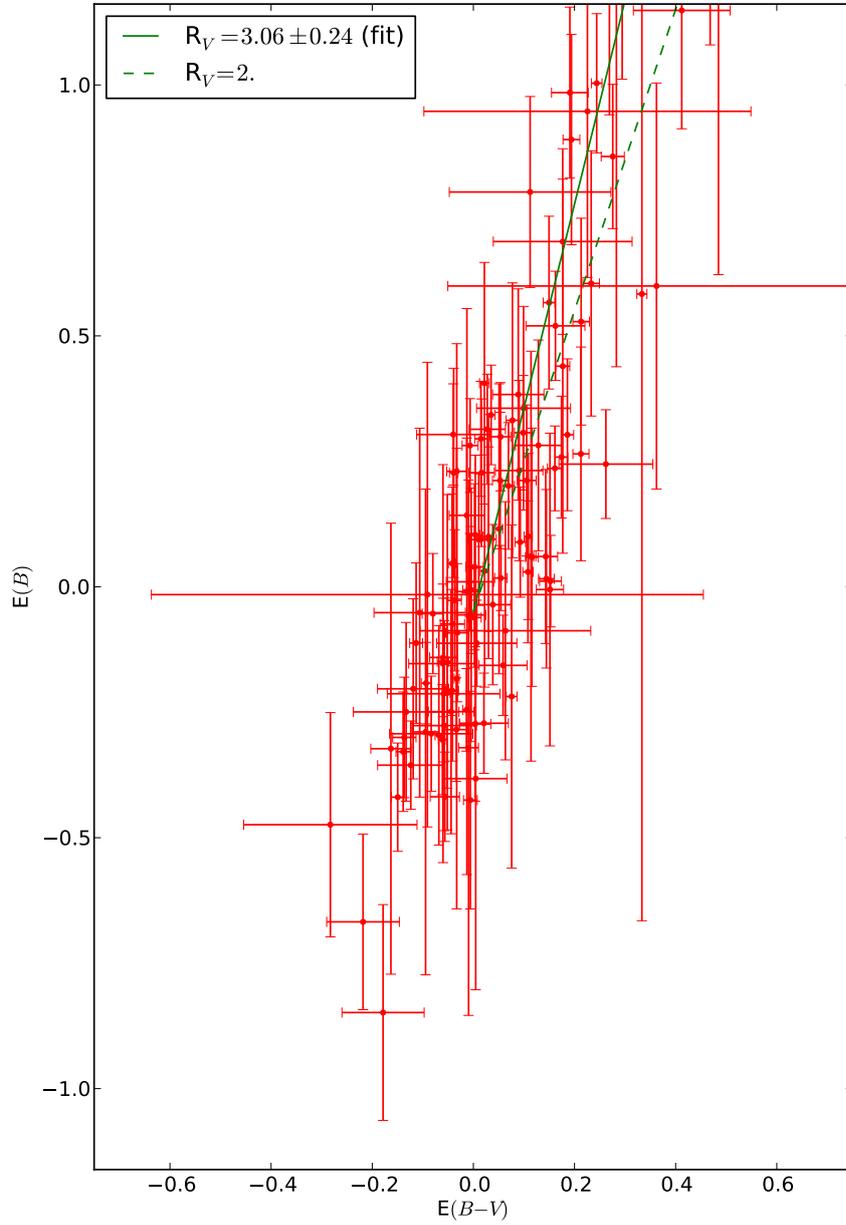


Figure 7.10: The plot shows the same relation as in Fig. 7.9, zooming in at the part with small reddening.

Chapter 8

Conclusions and Outlook

We have developed a new framework which allows the simultaneous characterization of large samples of spectra, forming an ideal ground for placing synthetic spectra among the observed ones. Combining Expectation Maximization Principal Component Analysis (EMPCA) and Partial Least Square (PLS) techniques, it defines a meaningful metric space and correlates it to spectroscopic and photometric intrinsic properties of supernovae.

The algorithm is based on the derivative of the spectrum over wavelength, which consequently assigns a larger weight to small scale features and, at the same time, makes the results independent of distance measurements, reddening and spectra calibration. The method allows an automatic exploration of information encoded in weak spectral features from the weak lines themselves, not only through their correlation with stronger lines. Moreover, the initial data matrix was forged to encode spectral evolution information through the use of spectral sequences representing each object. This shows an easy way to extract information from spectral evolution.

First, we applied the method to a large sample (~ 120 SNe and ~ 800 spectra) of well observed Type Ia supernovae obtained by the SNfactory collaboration and to a large sample of publicly available objects (~ 230 SNe and ~ 2100 spectra). We defined a low dimensional parameter space using EMPCA and studied the spectral features covered by each principal component separately. Results show that the high velocity features (HVF's) of Ca II H&K and infrared lines are uncorrelated with the properties of the rest of the ejecta, consistent with Mazzali et al. (2005). This suggests that the outer layers of the ejecta have variations partially unrelated to the inner structure. Surface layers do not allow to draw conclusions on the explosion mechanism.

We confirmed many of the results of Cormier & Davis (2011). For example, the properties of 91T-like SNe form a continuum with normal SNe, PCA can be used to form a continuum of spectral templates, and the first two PCs mainly describe the velocities of the ejecta and equivalent widths of the lines. A larger data set and the innovative method of analysing the derivative of the spectra allowed us to have a stable metric space without arbitrarily removing peculiar objects from the sample.

Once the PC space was defined, we applied the PLS algorithm in order to find directions in this low dimensional space which correlate with independently measured SNe Ia characteristics. In other words, we used the PC space as a tool which enables the reconstruction of not only the observed spectra, but also as a substitute of the spectral parameters often used to sub-classify SNe Ia. Among others, we investigated the velocity and pseudo-equivalent width (pEW) of Si II 5640 Å and Si II 6355 Å lines, the B and V magnitudes, the $B - V$ colors, the $\Delta m_{15}(B)$, and the SALT2 parameters c and x_1 . This demonstrates that the PC space is physically meaningful and includes the information recovered from commonly used spectral indicators. Moreover, it clarifies the potential of this framework to find missing or unexpected features in synthetic spectra. Our PLS results confirm the well known correlation between the pEW of Si II 5972 Å and the $\Delta m_{15}(B)$ in SNe Ia (Hachinger et al., 2006; Nugent et al., 1995).

The technique is not optimized to calibrate SN Ia. The observed color and magnitudes cannot be directly reconstructed by this technique alone, because they are largely contaminated by extinction. We show that the intrinsic $B - V$ color of SNe Ia is not constant among different objects and correlates with the velocity of Si II 6355 Å, as found also by Foley & Kasen (2011). We showed that the velocity of the Si II 5640 Å can be used for the same scope.

We build on the physically motivated assumption that the intrinsic light curve and color curve variability are a function of the variability in spectral series. We proved the applicability of the method to infrared light curves, both at maximum and few weeks after maximum. We showed that SN Ia do not have a uniform luminosity in the infrared, as thought by the majority of the community, but their luminosity correlates with the properties of spectra in the visible. On one hand, this proves that our metric space catches all the physically relevant variability of the bulk of SN Ia, on the other hand it shows that it is an excellent tool to calibrate SN Ia for cosmology using the more uniform infrared bands where dust extinction is largely irrelevant.

In the context of SN Ia, multivariate PLS, together with PCA, becomes a sharp tool able to separate the intrinsic variability and the variability due to dust. This approach can be valuable in a number of challenges, such as the study of the reddening law and its possible variability and improving the calibration of SNe Ia luminosity by the use of their spectra which, by definition, is a distance independent property. We started investigating the R_V dust parameter, the most commonly used parameter to characterize the type of dust. Our results offers a solution for the tension between the range of observed values for this parameter in the Milky-Way and the values inferred from SN Ia by the majority of the literature. The majority of the extinction of SNe Ia is caused by normal dust, similar to the typical Milky-Way dust. This suggests that most of the dust associated with SNe Ia is of interstellar origin and not associated with the progenitor. Our approach can easily be extended to other color bands and magnitudes in a broader wavelength range.

Now that we had the PCA trained on a large enough sample, we used our tool for the direct comparison between synthetic and real SN Ia spectra. Projecting a synthetic

spectral series in this PC space revealed its counterparts among the real data by an analysis of its neighbours. Moreover, the relations discovered by PLS are strong tests models have to pass. Given the challenge of performing a coherent statistical comparison between synthetic and real spectra, our method is particularly efficient in characterizing large sets of models built from different explosion scenarios. It is able to provide important insights regarding the global properties of each explosion mechanism in order to favour or disfavour them. Such a global analysis is also expected to be more robust against systematics in the models than comparing them individually on a case-by-case basis with real SNe.

Much of the known behaviour of the models is recovered in our work. For example, the pure deflagration models can be an explanation for faint SNe Ia. Similarly, also the merger models with equal initial masses ($0.9 - 0.9M_{\odot}$) are a good candidate for faint SNe Ia. The sub-Chandrasekhar mass detonations and the delayed detonations, on the other hand, are the best candidates for normal and bright SN Ia. In addition to what was known before, our tool offers very stringent tests for those models that are candidates for the bulk of SNe Ia, thanks to the abundant observational data. We found that the relations between spectroscopic and photometric properties predicted by the sub- M_{Chan} models do not follow the relations that are valid for the bulk of SN Ia. On the other hand, sub- M_{Chan} models with masses lower than $0.97M_{\odot}$ are excellent candidates for the faint SN Ia, where such relations are reproduced. Some of the shortcomings of the delayed detonation models as an explanation for the bulk of SNe Ia can be cured by a mechanism which reduces mixing in the brightest models. We proposed rotation as a possibility to achieve this result. A merger model with masses ($1.1 - 0.9M_{\odot}$), proposed as an explanation for the bulk of SNe Ia, evolves somewhat too slowly in relation to its spectral properties. This suggests an ejecta mass that is somewhat too high. On the other hand, it is positioned quite well in all of the other diagrams. An investigation of more merger models is necessary to reach definitive conclusions for this promising scenario.

Once a large enough library of synthetic spectra will be available, our method can also be used for the construction of a PC space based entirely on models and the projection of real objects in it, providing a cross-check between the real and synthetic metric spaces. A detailed study of such an application will be investigated in a future work.

Appendix A

A.1 Cross-Validation

We tested the stability of our PC space using a k -folding cross-validation (CV) algorithm. The goal of any CV procedure is to ensure that results are statistically consistent and not particular to a specific data set. At the same time, it tests for over-fitting. In our context, this means that even when applied to a sub-sample of the original data (training sample), the PC space configuration (Figure 4.5) must be recognizable. Moreover, the directions found by PLS in this space must be able to predict the values of the discrete observables for data not used in the EMPCA analysis (validation sample), using only their projections in PC space. Such results are expected to have residuals of the same magnitude for training and validation samples.

The number of foldings (k) denotes how the data will be divided between training and validation samples. The original set is divided into k mutually exclusive sub-samples and for each iteration one of these is stripped out of the original data set. The complete EMPCA and PLS algorithm is then applied to the remaining data and a linear fit is obtained characterizing the directions found by PLS and the discrete observables analysed in section 5. This process is repeated for all k subsamples and results for the PC projection and PLS analysis are stored in each iteration. The average displacement of each point in the PC space, calculated over all iterations, gives us a measurement of how much the stability of this space relies on individual data points. If the PC space configuration is highly unstable for different subsets, it can be considered evidence of the need of a larger, more representative, sample in order to safely draw conclusions. Analogously, an over-fitting method can be recognized if the PLS analysis is not able to provide estimations of the discrete observables for objects in the test sample, at least as accurately as it does for the training sample.

Here we present results for $k = 10$ foldings, which is a standard first choice for many CV procedures (Arlot & Celisse, 2010). However, we did perform the test for different values of k , with results following the expected behaviour: the PC space becomes more stable for larger values of k , the linear fits on the PLS results remain the same and the ratio of residuals between training and validation sets remain close to unity (Table A.1).

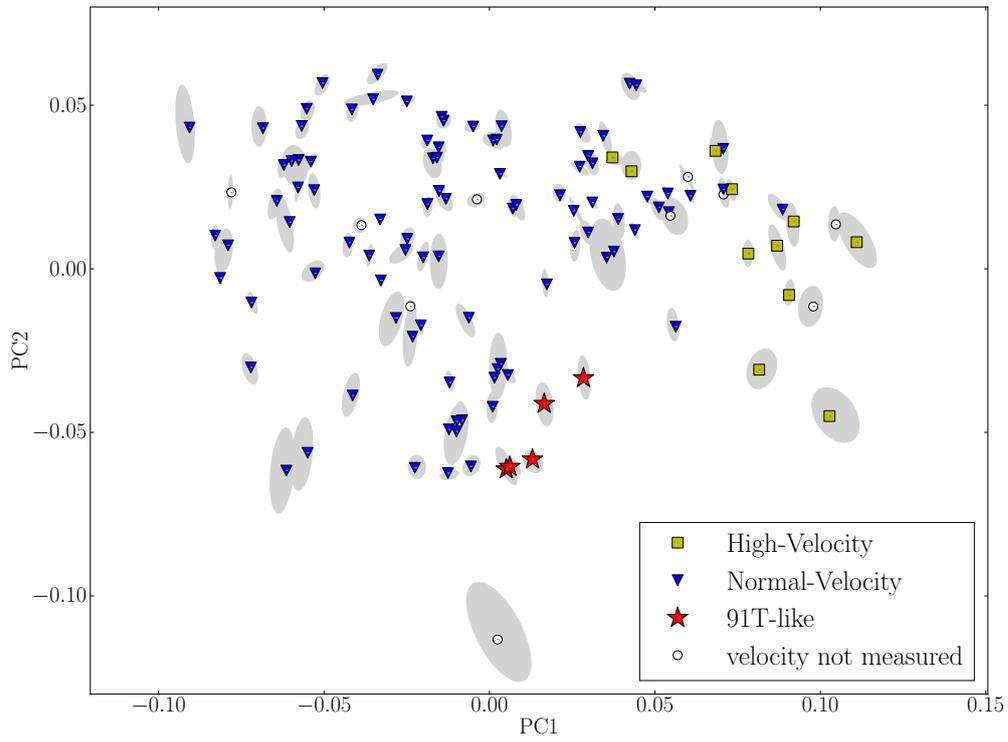


Figure A.1: Stability of PC space through $k = 10$ folding cross-validation. The color code for the points are the same used in Figure 4.3. The gray ellipses denote mean and 1σ variance for locations occupied by each data point throughout the 10 iterations.

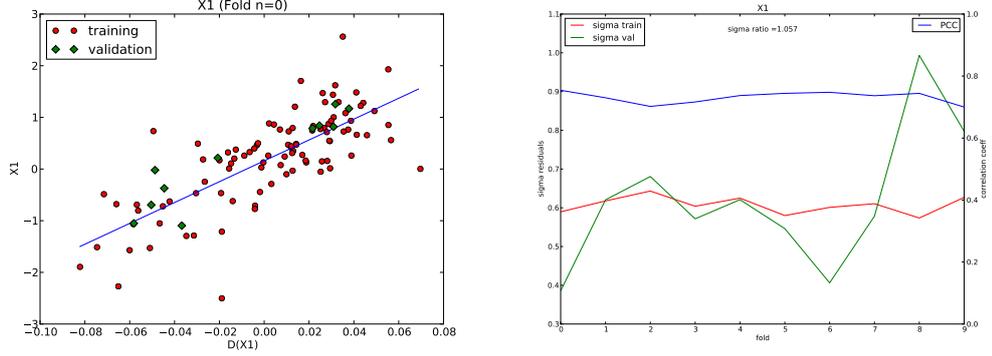


Figure A.2: Accuracy of PLS analysis in predicting the value of x_1 for the validation sample. **Left panel:** Results for one of the realizations. The red circles and green diamonds correspond to the training and validation sets respectively. The blue line shows the result from the linear fit applied to the training sample only. **Right panel:** Residuals from training (red) and validation (green) samples shown on the left axis, and Pearson correlation coefficient (PCC, in blue), shown in the right axis, for all 10 iterations. The average ratio between validation and training sample residuals is ≈ 1.057 .

The stability of the PC space in dF_{\log} is shown in Figures A.1 and A.2. The color code is the same used in Figure 4.3 and the gray ellipses represent the mean and 1σ variance of the locations occupied by each data point in all the 9 realizations in which it was part of the EMPCA. As an example, we show in the left panel of Figure A.2, the PLS results regarding the determination of x_1 , in one of the iterations. This plot illustrates how well the PLS is able to determine values of x_1 for points in the validation sample (green diamonds) in comparison with the variance present in the training sample (red circles). A more quantitative approach to such results throughout all the CV process is shown in the right panel of the same figure. Residuals from the determination of x_1 for training (red) and validation (green) samples, as well as the Pearson correlation coefficient (blue) for different folds (k) are shown. The mean ratio between residuals from validation and test samples was found to be very close to unity, verifying that our method is not suffering from over-fitting in the determination of discrete observables. Similar tests were performed for other observables and numerical results are shown in Table A.1.

	σ_{res} training	σ_{res} validation	ratio
Si II 6355–vel	608 km s ⁻¹	642 km s ⁻¹	1.06
S II 5640–vel	348 km s ⁻¹	362 km s ⁻¹	1.04
Si II 5972–pEW	5.5Å	6.0Å	1.09
Si II 6355–pEW	10.3Å	11.0Å	1.07
Δm_{15}	0.13	0.15	1.10
x_1	0.61	0.64	1.06

Table A.1: Residuals in estimation of observables from training and validation samples.

A.2 Line Velocities and Pseudo Equivalent Width Calculations

The values for line velocities and pEW used in section 5 were calculated using the algorithms described below.

In order to calculate the velocity of a line known to exist at an observed wavelength λ_0 , we start by searching for local minimum around λ_0 . Once the local minimum is found, we use its wavelength, the rest frame wavelength of the line and add relativistic corrections to compute the velocity blueshift.

If the line does not exist, the search for local minimum will lead us to the next important spectral feature and the final velocity value will be easy to recognize as wrong.

In computing the pEW, we need to determine the line tangent to the two nearest peaks surrounding a given spectral feature (A.2). We begin from the point of minimum flux of that feature (point A) and define two other points, along the flux function, to the left (point B) and to the right (point C) of point A. The area between the line connecting points B and C is calculated for successive small increments in the distances between A and B. The algorithm continues to iterate until the area between line BC and the flux function stop increasing. Once this maximum area is reached, B is kept fixed and the same procedure is applied to successive small increments in the distance between A and C. The calculation continues to alternate between increments in AB and AC until convergence. Once the maximum area is determined, it is used to characterize the pEW.

A.3 The reconstructions in the derivative space

In Figure C.1 we show the same reconstructions presented in Figure 4.8 in the original derivative space. We lack a physical intuition in observing this space and it is hard to recog-

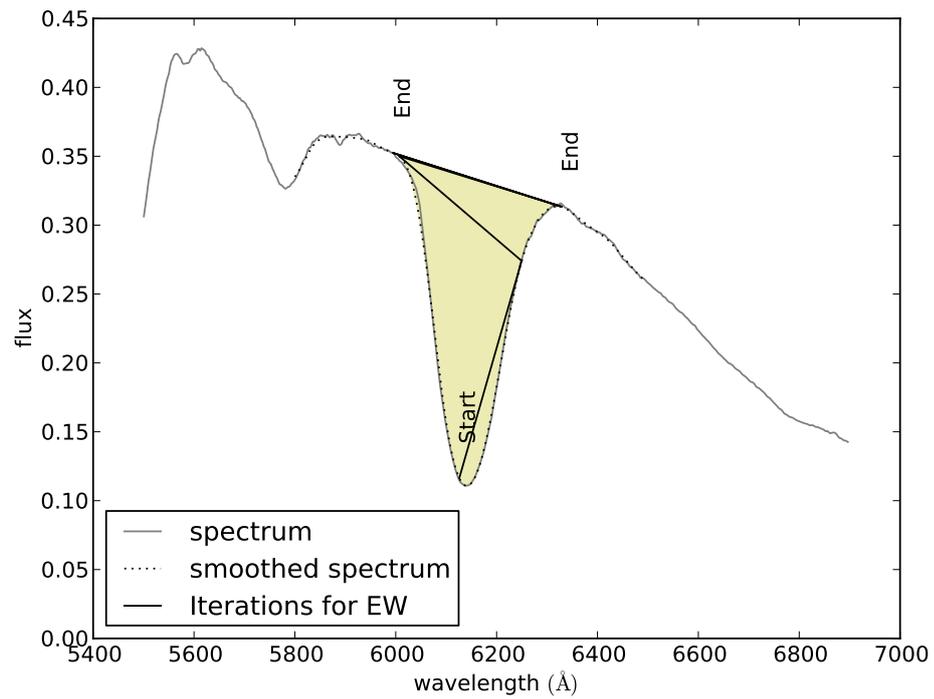


Figure A.3: To calculate the EW of lines we used a simple iterative algorithm shown in the figure.

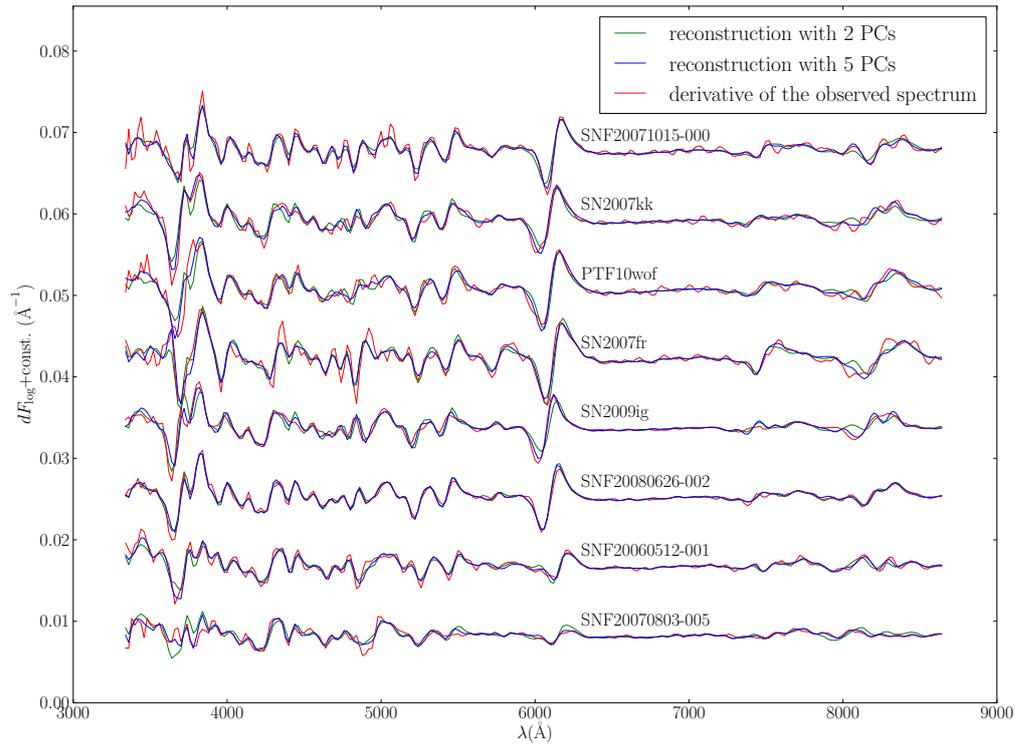


Figure A.4: Comparison between the derivative of the observed spectra (red) and reconstructions from PCA using 2 (green) and 5 (blue) PCs for a few supernovae at B -band maximum light.

nise the behaviour of the classical spectral indicators. However, it clearly demonstrates the ability of the derivative operation in minimizing reddening effects. It is instructive that the mismatches in color, which appear in the first two objects in Figure 4.8 (SNF20071015-000 and SN2007kk), are not noticeable anymore.

Bibliography

- Aldering G. et al., 2002, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 4836, Survey and Other Telescope Technologies and Discoveries, Tyson J. A., Wolff S., eds., pp. 61–72
- Arlot S., Celisse A., 2010, *Statistics Surveys*, 4, 40
- Arsenijevic V., Fabbro S., Mourão A. M., Rica da Silva A. J., 2008, *A&A*, 492, 535
- Aspden A. J., Bell J. B., Woosley S. E., 2010, *ApJ*, 710, 1654
- Bailer-Jones C. A. L., Irwin M., von Hippel T., 1998, *MNRAS*, 298, 361
- Bailey S., 2012, *PASP*, 124, 1015
- Bailey S., Aragon C., Romano R., Thomas R. C., Weaver B. A., Wong D., 2007, *ApJ*, 665, 1246
- Barbon R., Ciatti F., Rosino L., 1973, *A&A*, 25, 241
- Benetti S. et al., 2005, *ApJ*, 623, 1011
- Benitez-Herrera S., Ishida E. E. O., Maturi M., Hillebrandt W., Bartelmann M., Röpke F., 2013, *MNRAS*, 436, 854
- Benitez-Herrera S., Röpke F., Hillebrandt W., Mignone C., Bartelmann M., Weller J., 2012, *MNRAS*, 419, 513
- Bessell M. S., 1995, *PASP*, 107, 672
- Blondin S. et al., 2006, *AJ*, 131, 1648
- Blondin S. et al., 2012, *AJ*, 143, 126
- Boroson T. A., Green R. F., 1992, *ApJS*, 80, 109
- Branch D. et al., 2006, *PASP*, 118, 560
- Bronder T. J. et al., 2008, *A&A*, 477, 717

- Burns C. R. et al., 2014, *ApJ*, 789, 32
- Burns C. R. et al., 2011, *AJ*, 141, 19
- Buton C. et al., 2013, *A&A*, 549, A8
- Cardelli J. A., Clayton G. C., Mathis J. S., 1989, *ApJ*, 345, 245
- Charignon C., Chièze J.-P., 2013, *A&A*, 550, A105
- Childress M. et al., 2013, *ApJ*, 770, 107
- Chotard N. et al., 2011a, *A&A*, 529, L4
- Chotard N. et al., 2011b, *A&A*, 529, L4
- Conley A., Carlberg R. G., Guy J., Howell D. A., Jha S., Riess A. G., Sullivan M., 2007, *ApJL*, 664, L13
- Connolly A. J., Szalay A. S., Bershadsky M. A., Kinney A. L., Calzetti D., 1995, *AJ*, 110, 1071
- Contardo G., Leibundgut B., Vacca W. D., 2000, *A&A*, 359, 876
- Copin Y. et al., 2006, *NewAR*, 50, 436
- Cormier D., Davis T. M., 2011, *MNRAS*, 410, 2137
- Cox A. N., 2000, *Allen's astrophysical quantities*
- Dempster A. P., Laird N. M., Rubin D. B., 1977, *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39, 1
- Dhawan S., Leibundgut B., Spyromilio J., Maguire K., 2015, *MNRAS*, 448, 1345
- Diemer B., Kessler R., Graziani C., Jordan, IV G. C., Lamb D. Q., Long M., van Rossum D. R., 2013, *ApJ*, 773, 119
- Draine B. T., 2003, *ARA&A*, 41, 241
- Fink M. et al., 2014, *MNRAS*, 438, 1762
- Fink M., Röpke F. K., Hillebrandt W., Seitenzahl I. R., Sim S. A., Kromer M., 2010, *A&A*, 514, A53
- Folatelli G. et al., 2013, *ApJ*, 773, 53
- Foley R. J., Kasen D., 2011, *ApJ*, 729, 55
- Francis P. J., Hewett P. C., Foltz C. B., Chaffee F. H., 1992, *ApJ*, 398, 476
- Goobar A., 2008, *ApJL*, 686, L103

- Guy J. et al., 2007, *A&A*, 466, 11
- Hachinger S., Mazzali P. A., Benetti S., 2006, *MNRAS*, 370, 299
- Hawkins E. et al., 2003, *MNRAS*, 346, 78
- Hicken M. et al., 2009, *ApJ*, 700, 331
- Hillebrandt W., Kromer M., Röpke F. K., Ruiter A. J., 2013, *Frontiers of Physics*, 8, 116
- Hillebrandt W., Niemeyer J. C., 2000, *ARA&A*, 38, 191
- Hügelmeier S. D., Dreizler S., Werner K., Krzesinski J., Nitta A., Kleinman S. J., 2007, in *Astronomical Society of the Pacific Conference Series*, Vol. 372, 15th European Workshop on White Dwarfs, Napiwotzki R., Burleigh M. R., eds., p. 249
- Hunter J. D., 2007, *Computing In Science & Engineering*, 9, 90
- Iben, Jr. I., Tutukov A. V., 1984, *ApJS*, 54, 335
- Ishida E. E. O., de Souza R. S., 2011, *A&A*, 527, A49
- Ishida E. E. O., de Souza R. S., 2013, *MNRAS*, 430, 509
- Ishida E. E. O., de Souza R. S., Ferrara A., 2011, *MNRAS*, 418, 500
- James J. B., Davis T. M., Schmidt B. P., Kim A. G., 2006, *MNRAS*, 370, 933
- Jha S., Riess A. G., Kirshner R. P., 2007, *ApJ*, 659, 122
- Jolliffe I. T., 2002, *Principal Component Analysis*. Springer-Verlag
- Kaiser J., Hamming R., 1977, *Acoustics, Speech and Signal Processing*, *IEEE Transactions on*, 25, 415
- Kasen D., 2006, *ApJ*, 649, 939
- Kasen D., Thomas R. C., Nugent P., 2006, *ApJ*, 651, 366
- Kattner S. et al., 2012, *PASP*, 124, 114
- Krisciunas K. et al., 2011, *AJ*, 142, 74
- Kromer M. et al., 2013, *ApJL*, 778, L18
- Kromer M., Sim S. A., 2009, *MNRAS*, 398, 1809
- Kromer M., Sim S. A., Fink M., Röpke F. K., Seitenzahl I. R., Hillebrandt W., 2010, *ApJ*, 719, 1067

- Lantz B. et al., 2004, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 5249, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Mazuray L., Rogers P. J., Wartmann R., eds., pp. 146–155
- Ledoit O., Wolf M., 2004, *Journal of multivariate analysis*, 88, 365
- Mandel K. S., Narayan G., Kirshner R. P., 2011, *ApJ*, 731, 120
- Maund J. R. et al., 2013, *MNRAS*, 431, L102
- Mazzali P. A. et al., 2005, *ApJL*, 623, L37
- Mazzali P. A., Sauer D. N., Pastorello A., Benetti S., Hillebrandt W., 2008, *MNRAS*, 386, 1897
- Mohr P. J., Taylor B. N., Newell D. B., 2012, *Reviews of Modern Physics*, 84, 1527
- Moll R., Woosley S. E., 2013, *ApJ*, 774, 137
- Morrey J. R., 1968, *Analytical Chemistry*, 40, 905
- Nelson P. R., 2002, *Open Dissertations and Theses*, 1495
- Nelson P. R., MacGregor J. F., Taylor P. A., 2006, *Chemometrics and Intelligent Laboratory Systems*, 80, 1
- Nomoto K., 1982, *ApJ*, 253, 798
- Nomoto K., Thielemann F.-K., Yokoi K., 1984, *ApJ*, 286, 644
- Nugent P., Kim A., Perlmutter S., 2002, *PASP*, 114, 803
- Nugent P., Phillips M., Baron E., Branch D., Hauschildt P., 1995, *ApJL*, 455, L147
- Pakmor R., Kromer M., Röpke F. K., Sim S. A., Rüter A. J., Hillebrandt W., 2010, *Nat*, 463, 61
- Pakmor R., Kromer M., Taubenberger S., Sim S. A., Röpke F. K., Hillebrandt W., 2012, *ApJL*, 747, L10
- Pakmor R., Kromer M., Taubenberger S., Springel V., 2013, *ApJL*, 770, L8
- Pedregosa F. et al., 2011, *Journal of Machine Learning Research*, 12, 2825
- Pereira R. et al., 2013, *A&A*, 554, A27
- Pérez F., Granger B. E., 2007, *Computing in Science and Engineering*, 9, 21
- Phillips M. M., 1993, *ApJL*, 413, L105
- Phillips M. M., Lira P., Suntzeff N. B., Schommer R. A., Hamuy M., Maza J., 1999, *AJ*, 118, 1766

- Poludnenko A. Y., Gardiner T. A., Oran E. S., 2011, *Physical Review Letters*, 107, 054501
- Poznanski D., Nugent P. E., Filippenko A. V., 2010, *ApJ*, 721, 956
- Pskovskii I. P., 1977, *Soviet Ast.*, 21, 675
- Pskovskii Y. P., 1984, *Soviet Ast.*, 28, 658
- Rasmussen C. E., Williams C. K. I., 2005, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press
- Rau A. et al., 2009, *PASP*, 121, 1334
- Riess A. G., Press W. H., Kirshner R. P., 1996a, *ApJ*, 473, 88
- Riess A. G., Press W. H., Kirshner R. P., 1996b, *ApJ*, 473, 588
- Rigault M. et al., 2013, *A&A*, 560, A66
- Röpke F. K., 2007, *ApJ*, 668, 1103
- Röpke F. K. et al., 2012, *ApJL*, 750, L19
- Roweis S., 1998, *Advances in neural information processing systems*, 626
- Sasdelli M. et al., 2015, *MNRAS*, 447, 1247
- Sasdelli M., Mazzali P. A., Pian E., Nomoto K., Hachinger S., Cappellaro E., Benetti S., 2014, *MNRAS*, 445, 711
- Savitzky A., Golay M. J. E., 1964, *Analytical Chemistry*, 36, 1627
- Scalzo R. et al., 2012, *ApJ*, 757, 12
- Schlegel D. J., Finkbeiner D. P., Davis M., 1998, *ApJ*, 500, 525
- Schmidt W., Ciaraldi-Schoolmann F., Niemeyer J. C., Röpke F. K., Hillebrandt W., 2010, *ApJ*, 710, 1683
- Seitenzahl I. R. et al., 2013, *MNRAS*, 429, 1156
- Silverman J. M. et al., 2012, *MNRAS*, 425, 1789
- Sim S. A., Röpke F. K., Hillebrandt W., Kromer M., Pakmor R., Fink M., Ruiter A. J., Seitenzahl I. R., 2010, *ApJL*, 714, L52
- Sim S. A. et al., 2013, *MNRAS*, 436, 333
- Stehle M., Mazzali P. A., Benetti S., Hillebrandt W., 2005, *MNRAS*, 360, 1231
- Stritzinger M. D. et al., 2011, *AJ*, 142, 156

- Suzuki N., 2006, *ApJS*, 163, 110
- Tanaka M., Mazzali P. A., Maeda K., Nomoto K., 2006, *ApJ*, 645, 470
- Tripp R., 1998, *A&A*, 331, 815
- Wang B., Han Z., 2012, *NewAR*, 56, 122
- Wang X. et al., 2009, *ApJL*, 699, L139
- Webbink R. F., 1984, *ApJ*, 277, 355
- Weyant A., Wood-Vasey W. M., Allen L., Garnavich P. M., Jha S. W., Joyce R., Matheson T., 2014, *ApJ*, 784, 105
- Whelan J., Iben, Jr. I., 1973, *ApJ*, 186, 1007
- Whitney C. A., 1983, *A&AS*, 51, 443
- Wold H., 1982, *Systems Under Indirect Observation, PartII*, 36
- Wold S., Ruhe A., Wold H., Dunn, III W., 1984, *SIAM Journal on Scientific and Statistical Computing*, 5, 735
- Wold S., Sjöström M., Eriksson L., 2001, *Chemometrics and intelligent laboratory systems*, 58, 109
- Woosley S. E., 2007, *ApJ*, 668, 1109
- Yaron O., Gal-Yam A., 2012, *PASP*, 124, 668
- Yip C. W. et al., 2004, *AJ*, 128, 2603

Software references

SciPy – scipy.org – is an extensive open source library of scientific tools.

scikit-learn – scikit-learn.org – is a free-software tool for data mining and data analysis built on NumPy, SciPy and matplotlib (Pedregosa et al., 2011).

IPython – ipython.org – is a rich environment for interactive computing. Ideal for data analysis (Pérez & Granger, 2007).

EMPCA – github.com/sbailey/empca – Weighted Expectation Maximization Principal Component Analysis (Bailey, 2012).

matplotlib – matplotlib.org – is a 2D plotting library (Hunter, 2007).