Mon. Not. R. Astron. Soc. **000**, 1–19 (2014) Printed 18 November 2014 (MN LATEX style file v2.2)

A metric space for type Ia supernova spectra

Michele Sasdelli[1]∗, W. Hillebrandt[1], G. Aldering[2], P. Antilogus[3], C. Aragon[2], S. Bailey[2], C. Baltay[4], S. Benitez-Herrera[1], S. Bongard[3], C. Buton[5,6], A. Canto[3], F. Cellier-Holzem[3], J. Chen[7], M. Childress[2,8], N. Chotard[7,9], Y. Copin[10], H. K. Fakhouri[2,8], U. Feindt[5,6], M. Fink[1], M. Fleury[3], D. Fouchez[11], E. Gangler[10], J. Guy[3], E. E. O. Ishida[1,12], A. G. Kim[2], M. Kowalski[5,6], M. Kromer[1,13], S. Lombardo[5,6], P. A. Mazzali[14,1,15], J. Nordin[2,16], R. Pain[3], E. Pécontal[18], R. Pereira[10], S. Perlmutter[2,8], D. Rabinowitz[4], M. Rigault[5,6], K. Runge[2], C. Saunders[2], R. Scalzo[19], G. Smadja[10], N. Suzuki[2], C. Tao[7,11], S. Taubenberger[1], R. C. Thomas[17], A. Tilquin[11], B. A. Weaver[20]

[1] *Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Str. 1, 85741 Garching bei München, Germany*
[2] *Physics Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA*
[3] *Laboratoire de Physique Nucléaire et des Hautes Énergies, Université Pierre et Marie Curie Paris 6, Université Paris Diderot Paris 7, CNRS-IN2P3, 4 place Jussieu, 75252 Paris Cedex 05, France*
[4] *Department of Physics, Yale University, New Haven, CT 06520-8121, USA*
[5] *Physikalisches Institut, Universität Bonn, Nußallee 12, 53115 Bonn, Germany*
[6] *Institut fr Physik, Newtonstr. 15, 12489 Berlin, Humboldt-Universitt zu Berlin*
[7] *Tsinghua Center for Astrophysics, Tsinghua University, Beijing 100084, China*
[8] *Department of Physics, University of California Berkeley, 366 LeConte Hall MC 7300, Berkeley, CA, 94720-7300, USA*
[9] *National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China*
[10] *Université de Lyon, 69622, France; Université de Lyon 1, France; CNRS/IN2P3, Institut de Physique Nucléaire de Lyon, France*
[11] *Centre de Physique des Particules de Marseille, Aix Marseille Université, CNRS/IN2P3, CPPM UMR 7346, 13288 Marseille, France*
[12] *IAG, Universidade de São Paulo, Rua do Matão 1226, Cidade Universitária, CEP 05508-900, São Paulo, SP, Brazil*
[13] *The Oskar Klein Centre & Dept. of Astronomy, Stockholm University, AlbaNova, SE-106 91 Stockholm, Sweden*
[14] *Astrophysics Research Institute, Liverpool John Moores University, Liverpool L3 5RF, UK*
[15] *INAF-Osservatorio Astronomico di Padova, vicolo dell'Osservatorio, 5, I-35122 Padova, Italy*
[16] *Space Sciences Laboratory, University of California Berkeley, 7 Gauss Way, Berkeley, CA 94720, USA*
[17] *Computational Cosmology Center, Computational Research Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road MS 50B-4206, Berkeley, CA, 94720, USA*
[18] *Centre de Recherche Astronomique de Lyon, Université Lyon 1, 9 Avenue Charles André, 69561 Saint Genis Laval Cedex, France*
[19] *Research School of Astronomy and Astrophysics, The Australian National University, Mount Stromlo Observatory, Cotter Road, Weston Creek ACT 2611 Australia*
[20] *Center for Cosmology and Particle Physics, New York University, 4 Washington Place, New York, NY 10003, USA*

Accepted ... Received ...; in original form ...

ABSTRACT

We develop a new framework for use in exploring Type Ia Supernova (SN Ia) spectra. Combining Principal Component Analysis (PCA) and Partial Least Square analysis (PLS) we are able to establish correlations between the Principal Components (PCs) and spectroscopic/photometric SNe Ia features. The technique was applied to ~ 120 supernova and ~ 800 spectra from the Nearby Supernova Factory. The ability of PCA to group together SNe Ia with similar spectral features, already explored in previous studies, is greatly enhanced by two important modifications: (1) the initial data matrix is built using derivatives of spectra over the wavelength, which increases the weight of weak lines and discards extinction, and (2) we extract time evolution information through the use of entire spectral sequences concatenated in each line of the input data matrix. These allow us to define a stable PC parameter space which can be used to characterize synthetic SN Ia spectra by means of real SN features. Using PLS, we demonstrate that the information from important previously known spectral indicators (namely the pseudo-equivalent width (pEW) of Si II 5972 Å/Si II 6355 Å and the line velocity of S II 5640 Å/Si II 6355 Å) at a given epoch, is contained within the PC space and can be determined through a linear combination of the most important PCs. We also show that the PC space encompasses photometric features like B/V magnitudes, B-V colors and SALT2 parameters c and x_1. The observed colors and magnitudes, that are heavily affected by extinction, cannot be reconstructed using this technique alone. All the above mentioned applications allowed us to construct a metric space for comparing synthetic SN Ia spectra with observations.

Key words:
type Ia supernovae: general – Principal Component Analysis, derivative spectroscopy, Partial Least Square analysis, reddening and intrinsic color

1 INTRODUCTION

Type Ia Supernovae (SNe Ia) are among the most luminous transients in the Universe. They appear to be a rather homo-

∗ E-mail: sasdelli@mpa-garching.mpg.de

geneous group, both photometrically and spectroscopically. After the discovery of a relation between their light-curve shape and luminosity (Phillips 1993) 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. This increases the need to identify their progenitors and to understand the explosion mechanism. It is widely accepted that SNe Ia are the result of the thermonuclear explosion 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).

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). Studying their spectral differences is a promising way to shed some light on questions regarding their nature. 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 Objects[1] (PESSTO, e.g. Maund et al. 2013) and a large number 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 SUSPECT [2] 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.

In order to explore this potential, 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* was 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 this work, we 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 was 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 was obtained, we used 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 as SALT2 (Guy et al. 2007) parameters c and x_1. The outcomes from this analysis, applied to data from SNfactory, enabled the construction of a metric space where any given synthetic spectrum can be projected and automatically confronted with real data ones. Systematic comparisons of models with observation have been 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.

The paper is organized as follows: in Section 2 we present details about all pre-processing techniques and statistical methods used to build 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. Section 3 describes the SNfactory data set and the additional spectroscopic and photometric features to be investigated through PLS algorithm. Results from the EMPCA analysis (Bailey 2012), including illustrative comparisons to models, are presented in Section 4. Section 5 presents the independently measured SNe Ia features investigated in this work and the corresponding results from PLS are shown in Section 5.1. Finally, our conclusions are delineated in Section 6.

2 METHOD

2.1 Weighted Savitzky-Golay filter

Before attempting any process of information extraction on spectral data, one must take into account the high impact of random noise originated 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.

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

[1] `http://www.pessto.org`
[2] `http://www.nhn.ou.edu/~suspect`

(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 routine[3], where the quantity to be minimized is given by

$$S = \sum_{i=1}^{N} \left[w_i \left(F_{\lambda_i}^{\rm obs} - g_{\rm M}(\lambda_i, \boldsymbol{\beta}) \right) \right]^2 . \quad (1)$$

Here, N is the number of data points included in a fixed window, $F_{\lambda_i}^{\rm obs}$ is the observed flux at wavelength λ_i, $g_{\rm M}$ is the polynomial of degree M, $\boldsymbol{\beta}$ is the vector of scalar coefficients of g and w_i is the weight assigned to $F_{\lambda_i}^{\rm obs}$. 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. Although other types of smoothing techniques might possibly improve the results of our analysis, this matter has not been investigated in detail in this work.

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 PCA

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,

$$\Sigma_{ii'} = \frac{\sum_{k=1}^{k=N} (X_i^k - \overline{X_i})(X_{i'}^k - \overline{X_{i'}})}{N}, \quad (2)$$

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 Jollife 2002). Hereafter, we will always refer to the initial data as the mean subtracted terms in Eq. 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 ($M \ll N$),

$$\boldsymbol{F}_{\rm rec} \approx \overline{\boldsymbol{X}} + \sum_{j=1}^{M} c_j \boldsymbol{P}_j, \quad (3)$$

with $\overline{\boldsymbol{X}}$ representing the mean of all spectra, $\boldsymbol{P}_j$ the j-th PC and c_j the j-th scalar whose values must be determined from fitting $\boldsymbol{F}_{\rm rec}$ to the measured flux. Geometrically, c_j represents the projection of the measured spectrum on $\boldsymbol{P}_j$. 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 3, we can think of the PCs as the vectors which minimize $\chi^2 = \sum_{k=1}^{N} \left[\boldsymbol{X}^k - \boldsymbol{F}_{\rm rec} \right]^2$. 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,

$$\chi^2 = \sum_{k=1}^{N} \mathbf{W}^k \left[\boldsymbol{X}^k - \boldsymbol{F}_{\rm rec} \right]^2 . \quad (4)$$

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).:

Algorithm 1 Expectation Maximization algorithm

(i) $\mathbf{V} \leftarrow$ random orthonormal basis of dimension $i \times M$
(ii) repeat until convergence (i.e. the basis 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 V which minimizes equation 4 (M-step)
  (c) normalize the columns of V to unit length
(iii) return $\mathbf{V}$ as the EMPCA calculation of the first M eigenvectors of the basis $\mathbf{P}$

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

[3] `http://docs.scipy.org/doc/numpy/reference/generated/numpy.polyfit.html`

and EMPCA, we are able to 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 (Jollife 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, the variance in 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 in 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 was submitted to the SG filter and a random noise based on the original error amplitude was added to the smoothed spectrum. The new noisified spectrum was again submitted to the filtering process and its corresponding projection in PC space was computed. The procedure was 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:

$$F_{\log} = \log_{10} F_\lambda^{\rm obs} = \log_{10} F_\lambda^{\rm intr} - 0.4 \frac{A_\lambda}{A_V} R_V E_{B-V}, \quad (5)$$

where $F_\lambda^{\rm obs}$, $F_\lambda^{\rm intr}$ and A_λ are the observed flux, intrinsic flux and extinction at wavelength λ, respectively. A_V represents the extinction in 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 to 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 ac-

counts 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 null 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 subsection we show how 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 $\boldsymbol{e}$ ($\sum_i e_i^2 = 1$) that maximizes

$$Cov(\boldsymbol{e}X, Y) = \frac{\sum_{k=1}^{N} (Y^k - \overline{Y}) \sum_i (X_i^k - \overline{X_i}) e_i}{N}, \qquad (6)$$

where N is the number of objects and $\overline{X_j}$ and $\overline{Y}$ are means:

$$\overline{X_j} = \frac{\sum_{k=1}^{N} X_j^k}{N}, \; \overline{Y} = \frac{\sum_{k=1}^{N} Y^k}{N}.$$

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

$$[\sigma(\sum_i X_i e_i)]^2 = \frac{\sum_{k=1}^{N} (\sum_i (X_i^k - \overline{X_i}) e_i)^2}{N},$$

$$[\sigma(Y)]^2 = \frac{\sum_{k=1}^{N} (Y^k - \overline{Y})}{N}.$$

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.

3 APPLICATION

In this section we apply the previously explained framework to SN Ia spectra from the SNfactory.

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 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).

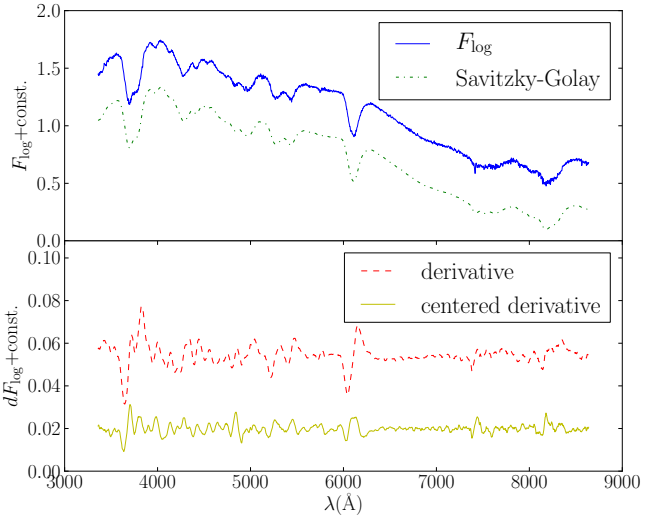

Figure 1. Multiple steps in data treatment. Both panels show data from SNF20080626-002, taken at -0.65 days relative to $B-$band maximum brightness. In each panel we artificially shifted the curves along the vertical axis for didactic reasons. **Top**: $F_{\log}$ measurements before (blue-full) and after (green-dotted) going through the SG filtering. **Bottom**: $dF_{\log}$ (red-dashed) and center derivative, $dF_{\log} - \overline{dF_{\log}}$ (yellow-full).

3.2 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. The $\Delta m_{15}(B)$ of the sample is within 0.7 and 1.7, the SALT2 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. (2011) 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 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 SNf 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

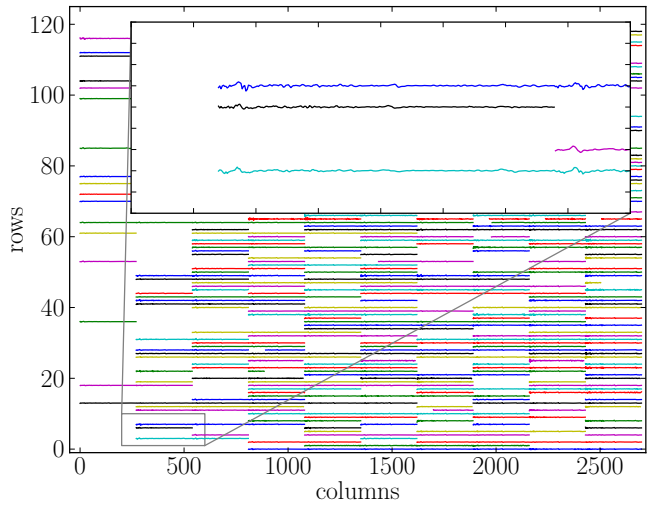

Figure 2. Representation of the input data matrix. Different rows correspond to different SNe. Each column shows centered $dF_{\log}$, from spectra collected between -10 and $+10$ days relative to $B-$band maximum brightness (from left to right), within each 2 day epoch window. Each curve runs over 3300Å$\leqslant \lambda \leqslant$ 9000 Å, written in wavelength bins of 20Å.

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 were 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 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 2 illustrates the overall shape of the final data matrix. Each spectrum was sampled every 20Å (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

$$F_\lambda^{\rm obs} \rightarrow g_3(\lambda,\overline{\beta}) = \beta_0+\beta_1(\lambda-\lambda_0)+\beta_2(\lambda-\lambda_0)^2+\beta_3(\lambda-\lambda_0)^3, \tag{7}$$

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:

$$F_{\log} = \log_{10} F_\lambda^{\rm obs}\Big|_{\lambda=\lambda_0} = \log_{10}\beta_0, \tag{8}$$

$$dF_{\log} = \frac{d\log_{10} F_\lambda^{\rm obs}}{d\lambda}\Big|_{\lambda=\lambda_0} = \frac{\beta_1}{\beta_0\ln 10}, \tag{9}$$

finally, propagating the errors:

$$\delta F_{\log} = \frac{\delta\beta_0}{\beta_0\ln 10}, \tag{10}$$

$$\delta dF_{\log} = \left|\frac{\beta_1}{\beta_0\ln 10}\right|\sqrt{\frac{\delta\beta_0^2}{\beta_0^2}+\frac{\delta\beta_1^2}{\beta_1^2}-\frac{2{\rm cov}(\beta_0,\beta_1)}{\beta_0\beta_1}}, \tag{11}$$

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 ${\rm cov}(\beta_0,\beta_1)$. 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.

4 PRINCIPAL COMPONENTS INTERPRETATION AND METRIC SPACE COMPARISON

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 3 and 4, and Figures 5 and 7). Hereafter, the PCs derived from a data matrix based on $F_{\log}$ will be referred to as PC$i^{F_{\log}}$, with i denoting the PC number. Meanwhile, PC 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 3 and 4 show the behaviour (first panel) and consequent influence on 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 3) 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 is also carries some discrete wavelength dependent features. Also, it clearly describes a much larger variance than the next two components (larger area covered by the gray region, second panel of Figure 3). The first PC is largely influenced by dust, with its long wavelength behaviour being consistent with a Cardelli reddening law. 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, 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 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, which tracks the variation of the High Velocity Features (HVFs) 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 is a property of the outer layers of the ejecta and it is 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 possibly be recognized also in synthetic spectra.

4.3 Metric Spaces

The projection of SNfactory data in a 2-dimensional PC space, obtained from $F_{\log}$, is displayed in Figure 5. 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 5 can be considered to be an alternative visualization of the same effect as presented in Figure 3: 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 $\mathrm{E}(B-V) = 0.4$ with a Cardelli et al. (1989) law (magenta line in Figure 5). This object has an observed $B-V$ color of ~ 0.4, one of the reddest SNe in the SNfactory sample. Figure 6 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 7 shows how this situation changes when the analysis is based on $dF_{\log}$. The crosses due to the instability of the EMPCA algorithm are completely negligible, the ellipses due to noise and missing components are large only in 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

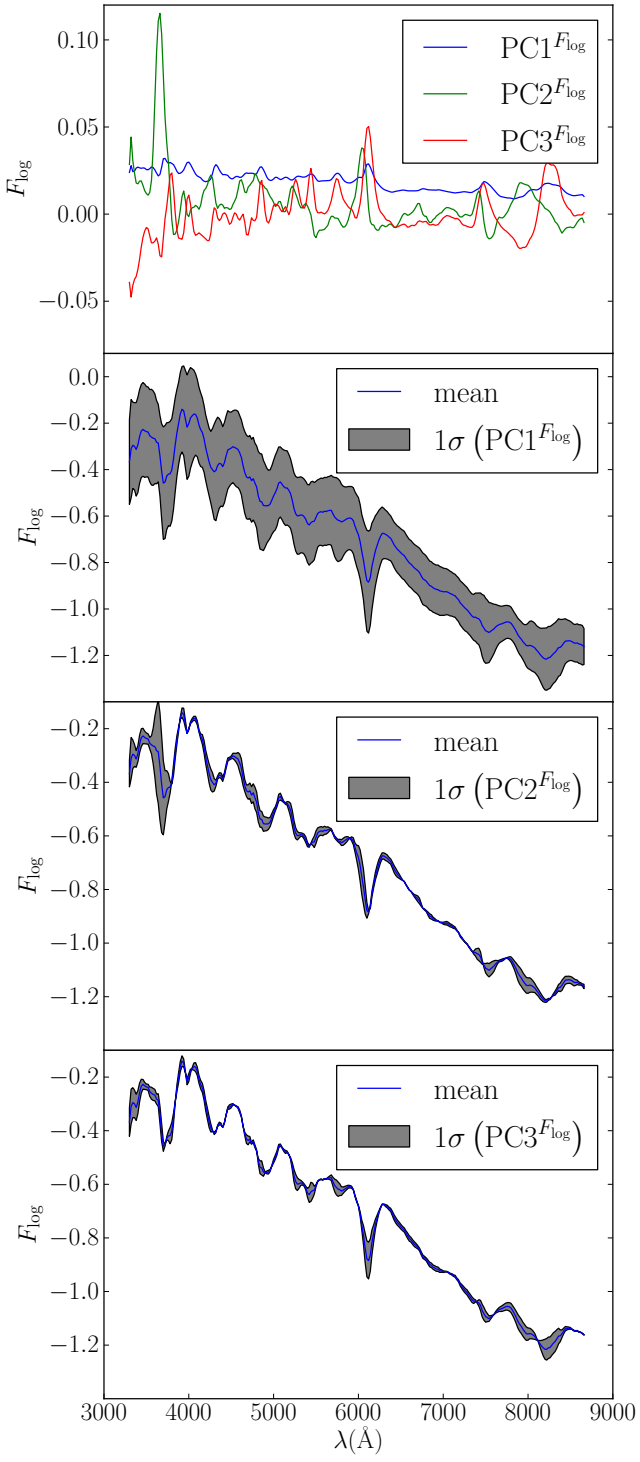

Figure 3. First panel shows the first three eigenvectors obtained from the analysis on $F_{\log}$. The second to fourth panels illustrate the main spectral features tracked by PC1, PC2 and PC3. All panels correspond to a spectrum taken between −6 and −4 days relative to B-band maximum. Blue lines denote 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).

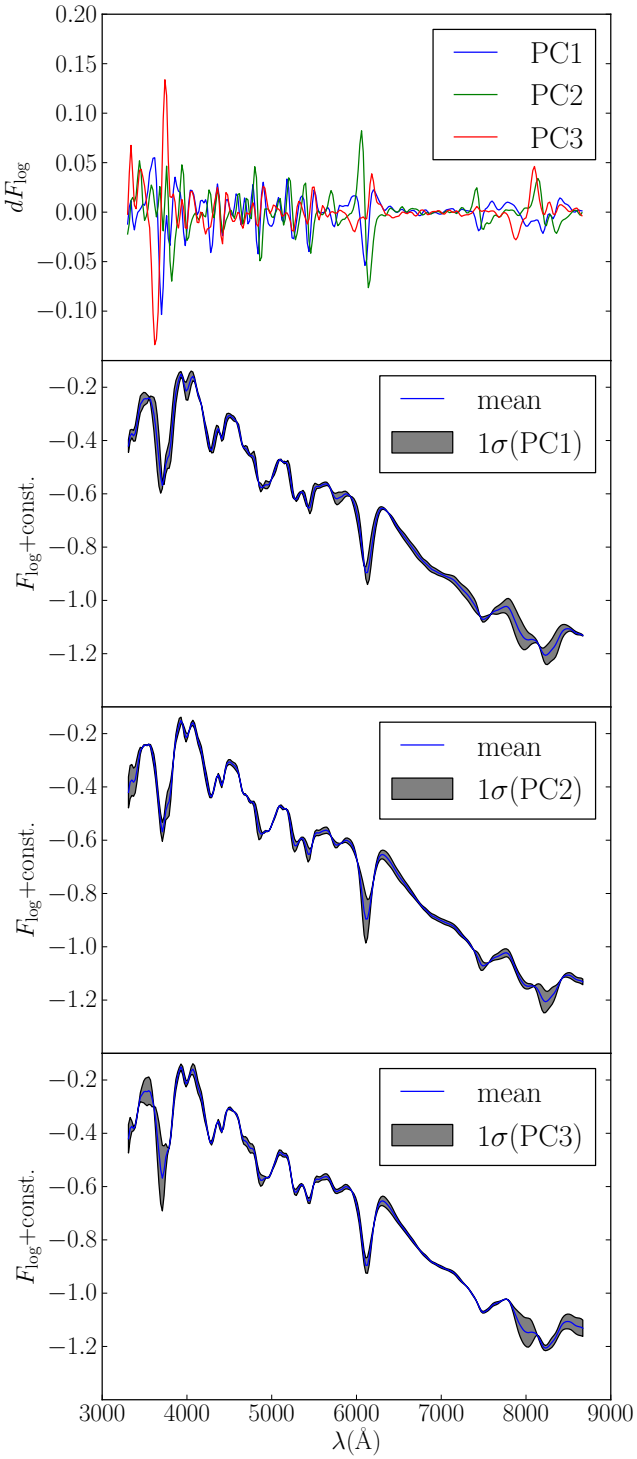

Figure 4. Same as Figure 3, but from a data matrix based on $dF_{\log}$.

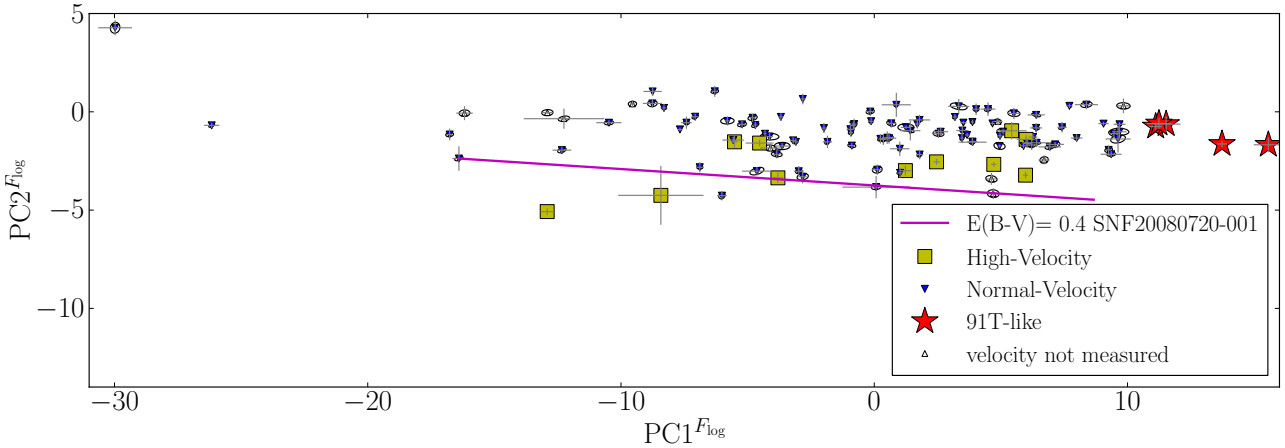

Figure 5. Projections of SNfactory data on the first two PCs for analysis based on $F_{\log}$. Each point represents a supernova, colored according to the spectral classification of Wang et al. (2009). A few 91T-like SNe are also highlighted. The crosses correspond to 1σ errors coming from random seed 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.

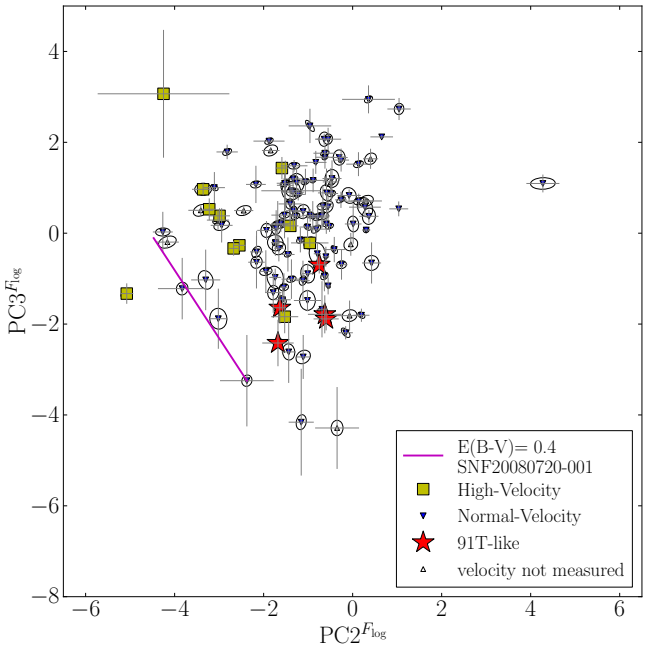

Figure 6. Same as Figure 5, PC2$^{F_{\log}}$ and PC3$^{F_{\log}}$.

(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 7. As in Figure 5, 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 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 2009). Detailed results from a CV test are presented in Appendix A, and these demonstrate the stability of the space presented in Figure 7.

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

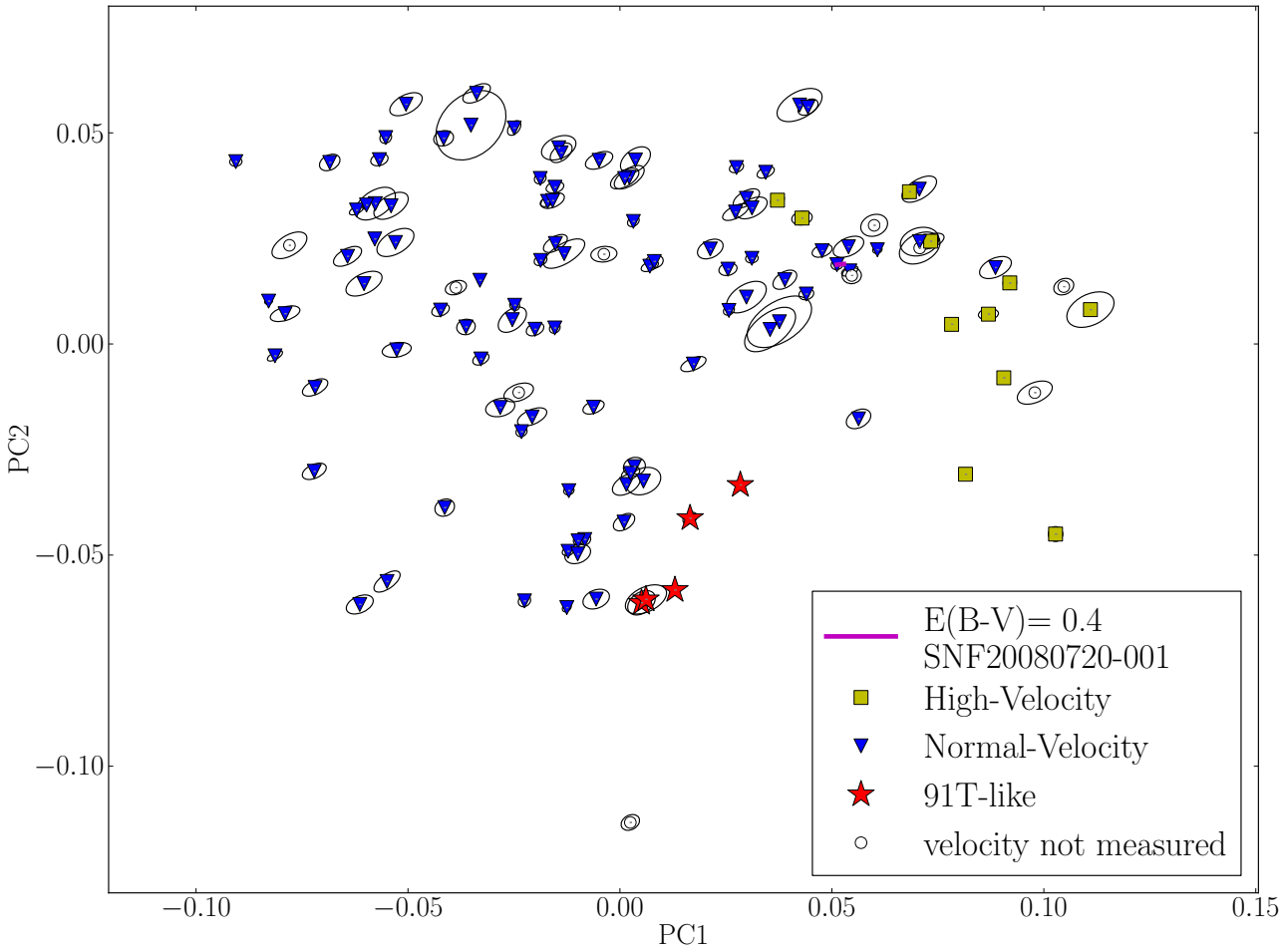

Figure 7. Same as Figure 5, for the analysis based on $dF_{\log}$.

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 (Jollife 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 8 and 9 we show the variances normalized to the first component for the analysis on $F_{\log}$ and $dF_{\log}$ respectively. Figure 8 shows the same result we have seen in Figures 3 and 5, with most of the information concentrated in $\mathrm{PC1}^{F_{\log}}$. 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 5 and 6). 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 9, 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 10 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

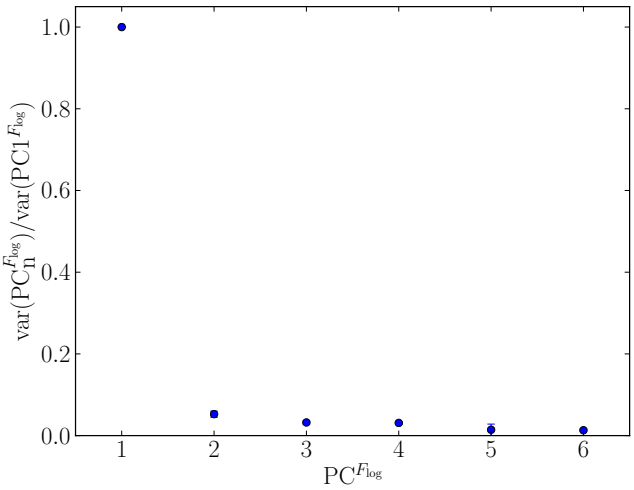

Figure 8. Distribution of variance among the first 6 PCs from $F_{\log}$ data matrix. The variances are normalized to that of the first PC. The errorbars show the variability due to k-folding (Appendix A).

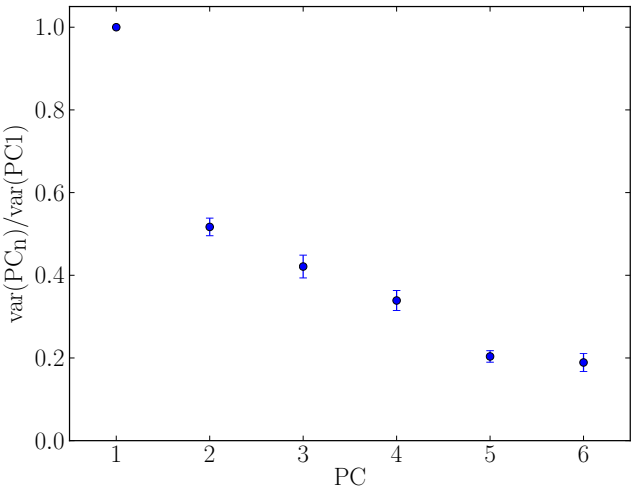

Figure 9. Same as Figure 8, but obtained from $dF_{\log}$ data matrix.

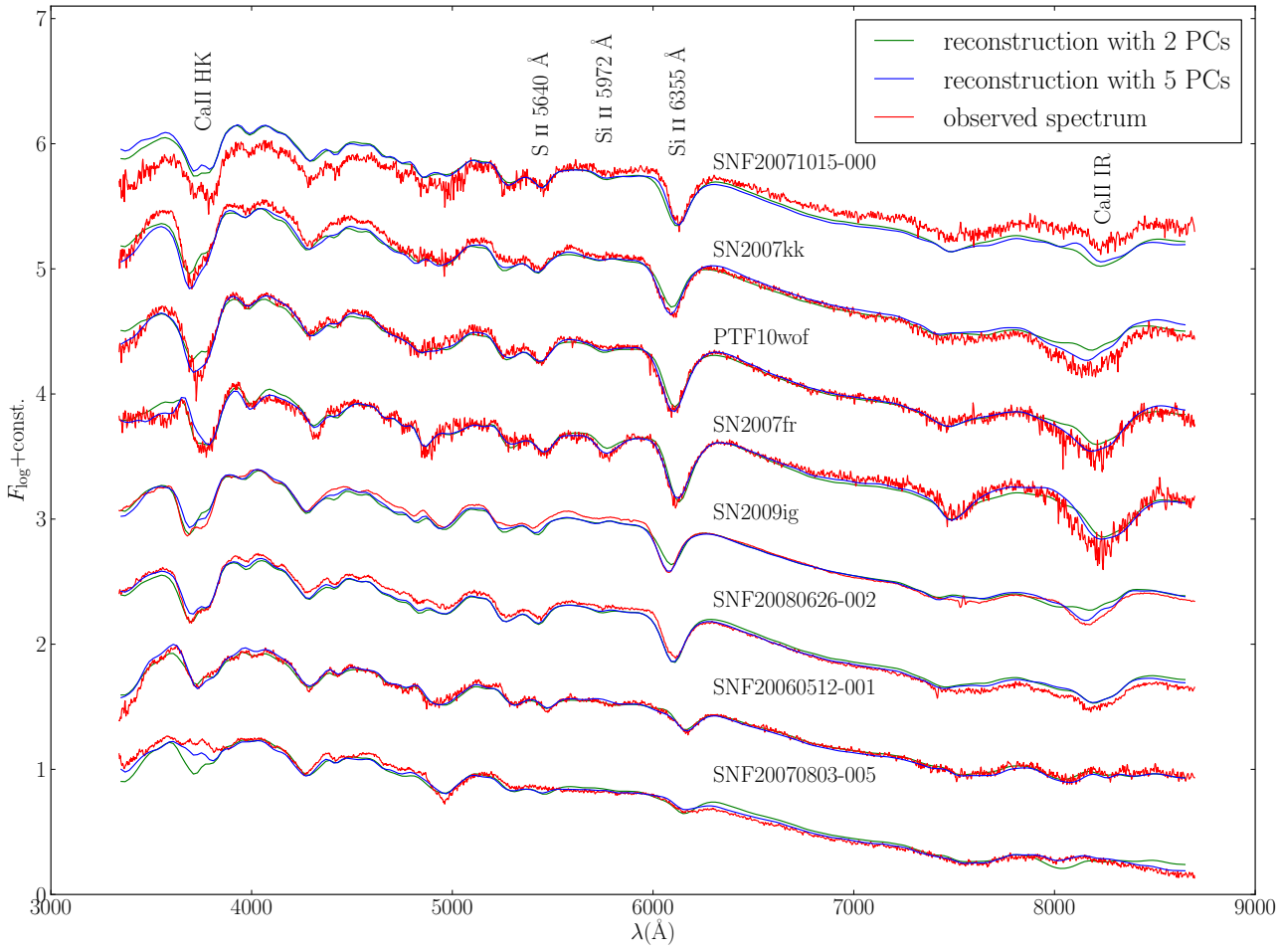

Figure 10. Comparison between the observed spectra without smoothing (red) and reconstructed spectra using 2 (green) and 5(blue) PCs, in the $dF_{\log}$ approach, for a few supernovae at $B-$band maximum light.

SNe at maximum for clarity, but this behaviour holds for all epochs between −10 to +10 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 C.

4.4 Comparison with models

We emphasis the potential of this metric space, built from the combination of the derivative approach and EMPCA, in providing an ideal environment for the characterization of synthetic spectral series (from −10 to +10 days) within a space defined by observed SNe. As an example, the black symbols in Figure 11 denote the projection, in derivative PC space, of the 3D delayed detonation model "N100" (Seitenzahl et al. 2013; Sim et al. 2013) and the merger model from Pakmor et al. (2012).

The N100 model describes a supernova generated from a white dwarf accreting material from a companion and getting close to the Chandrasekar mass (1.4 $M_\odot$). The other model describes the explosion of two carbon-oxygen white dwarfs with masses of $0.9M_\odot$ and $1.1M_\odot$ prototypical for the double degenerate scenario. The model spectra were constructed based on radiative transfer calculations of Kromer & Sim (2009) and projected into the PC space through the same procedure applied to the observed sample. Both models have been proposed as explanation of normal SN Ia, in particular SN2011fe (e.g. Röpke et al. 2012), and have a similar luminosity.

Figure 11 demonstrates that our procedure does place both models among the normal SNe (blue dots), in derivative PC space. However, the geometrical distance between them is considerable, reflecting the intrinsic and spectral differences of these two models (Röpke et al. 2012). This illustrates the power of our method when applied to characterize SN Ia models, providing an automatic and quantitative approach to confront them with observations and with each other. Such investigation will be further developed in a subsequent work.

5 COMPARISON WITH DISCRETE OBSERVABLES

In the context of the PLS we will now study the correlation 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 aimed at properly standardizing them. The empirical relation between brightness and decline rate demonstrated by (Phillips 1993) is considered one of the first standardization techniques for SNe Ia. It is given in terms of Δm15(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 Δm15(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 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) 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. 2011), 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, lightcurve 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 Δm15(B) with the fewest possible modeling assumptions. Wherefore, 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 Δm15(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

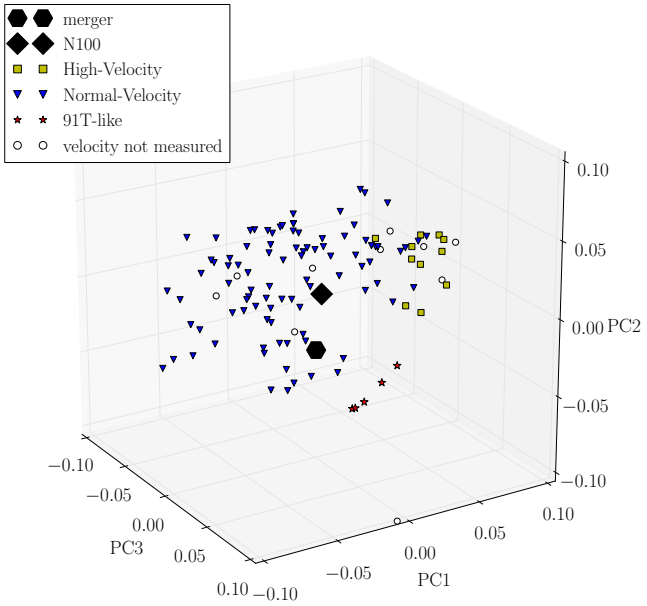

Figure 11. The merger (star) and the N100 delayed detonation (diamond) models, projected in to the first 3 components in derivative space.

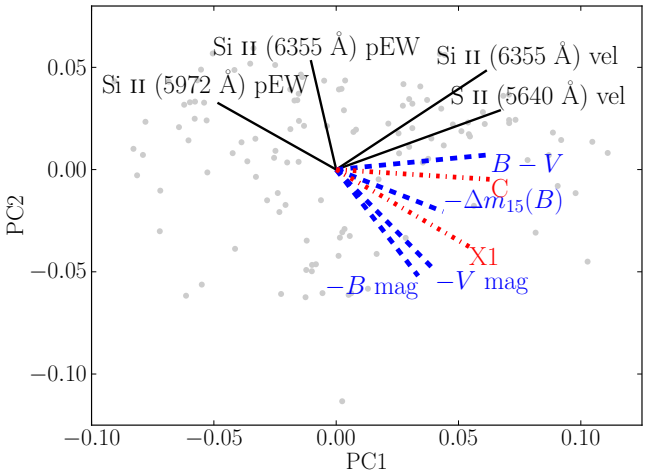

Figure 12. The directions maximizing the covariance with various SN parameters derived in a 5-dimensional space and projected into the plane formed by the first two principal components. Gray points are the same as those shown in Figure 7. Directions correlated with spectroscopic quantities are coloured in black (solid), photometric quantities in blue (dashed), and results from the SALT2 fit in red (dash-dotted).

and added in quadrature to uncertainties due to peculiar velocity of the host galaxies of $\sim$ 500 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 Δm15(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).

	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(Bmag)	0.40	−0.63	0.49	−0.32	−0.32
D(Vmag)	0.49	−0.60	0.49	−0.21	−0.34
D($B-V$)	0.76	0.09	0.43	0.45	−0.13
D(c)	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 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).

	Pearson coeff.	$\sigma_{\rm res}$
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 2. Pearson correlation coefficient for the linear fit between the directions found by PLS and independently measured observables. $\sigma_{\rm res}$ corresponds to the mean residual between the measured observables values and those determined through PLS.

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 B. 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 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 PLS

In Section 3.2, 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 connections 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 Section 4.3 (1D - $\mathcal{Y}$).

Figure 12 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 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 Åand Δm15(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 that of Δm15(B) (i.e. opposite to $-\Delta$m15(B), Figure 12). 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 the best with indicators of color.

In Table 2 we present the correlations given by PLS for SNe features with each one of the directions highlighted in Figure 12. The fact that many important SN features have strong signatures in our new metric spaces gives us confidence that our framework can help us better place synthetic spectra among their real data counterparts. Next we examine these trends in more detail.

5.3 Spectroscopic observables in derivative PC space

Figure 13 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 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 4.1). The few outliers on the high-velocity side of Figure 13 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.

Our ability to describe the velocity of S II 5640 Å using the 5D PC space is illustrated in Figure 14. 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 possible a better measure of the velocity. More generally, S II lines are not affected by HV features, which can complicate the measurement 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 for 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 study 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. 13 to 15 are only valid at maximum. An analogous study aimed at a different epoch would require the determination of spectral features at the epoch in question.

Figures 15 and 16 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 PC 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 Δm15(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 and dust contributions.

Figures 17 and 18 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 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 extincted SNe Ia, under the assumption that brighter SNe Ia have less extinction, using an iterative rejection scheme. This type of

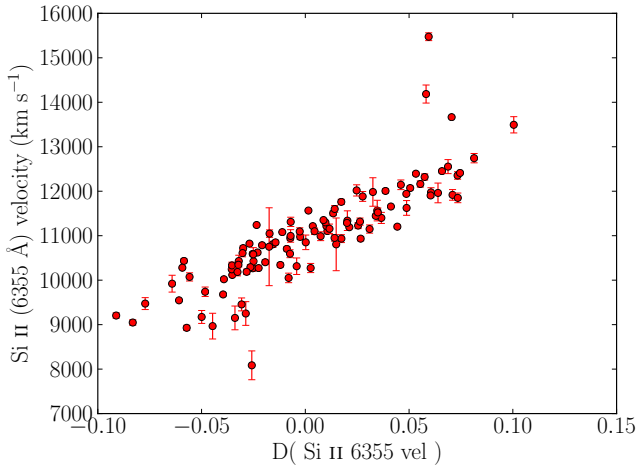

Figure 13. Correlation between PLS result and the Si II 6355 Å velocity at B-band maximum. The few outliers on the high-velocity side are due to HVFs of Si (see text).

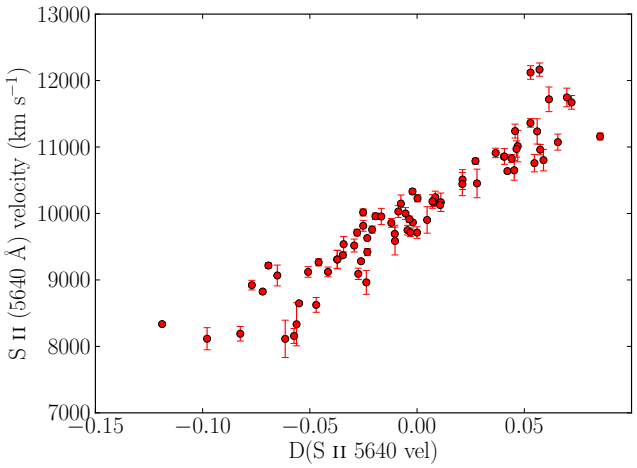

Figure 14. Same as Figure 13, but for the S II 5640 Å velocity at B-band maximum.

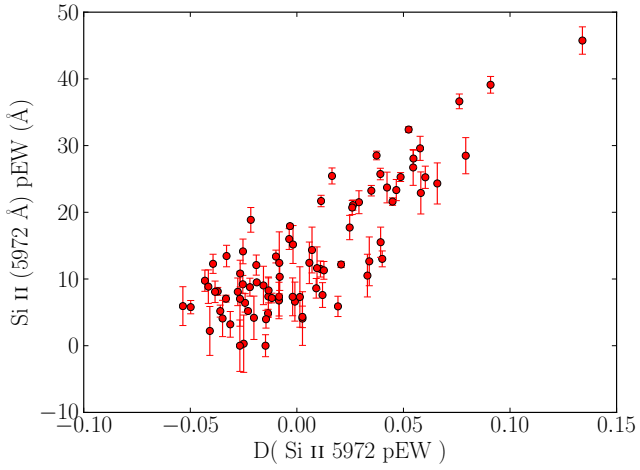

Figure 15. Same as Figure 13, but for the Si II 5972 Å pEW at B-band maximum.

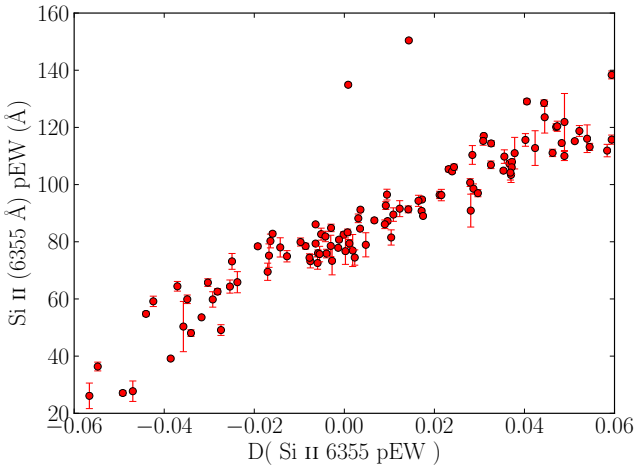

Figure 16. Same as Figure 13, but for the Si II 6355 Å pEW.

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 supernova brighter than the linear fit, or fainter by less than 0.3 mag, are selected to 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 17 and 18 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 19 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 20 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

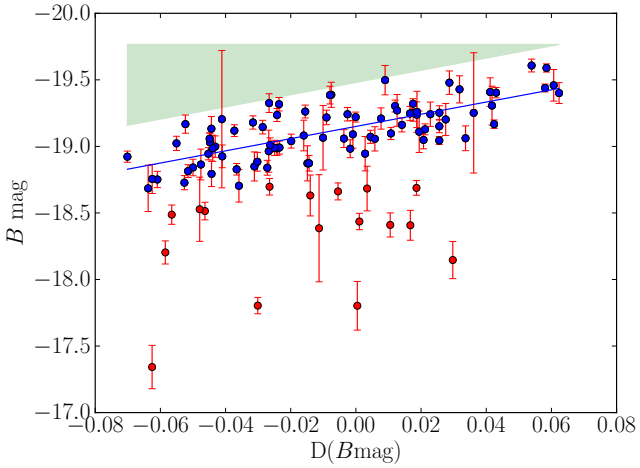

Figure 17. Correlation between PLS result and the B-band magnitude. The red points belong to SNe much redder than others with the same spectral characteristics.

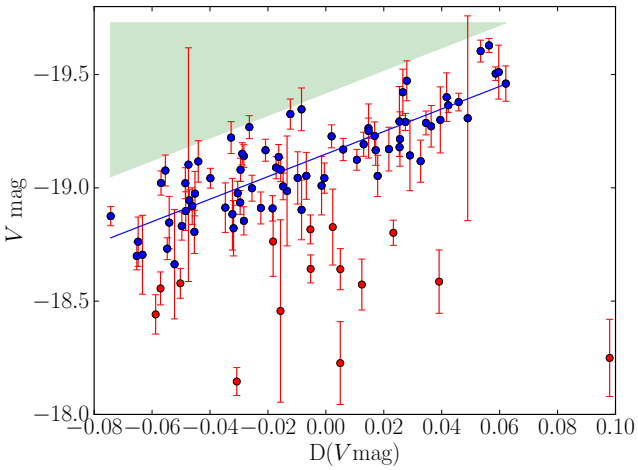

Figure 18. Same as Figure 17 for V-band magnitude.

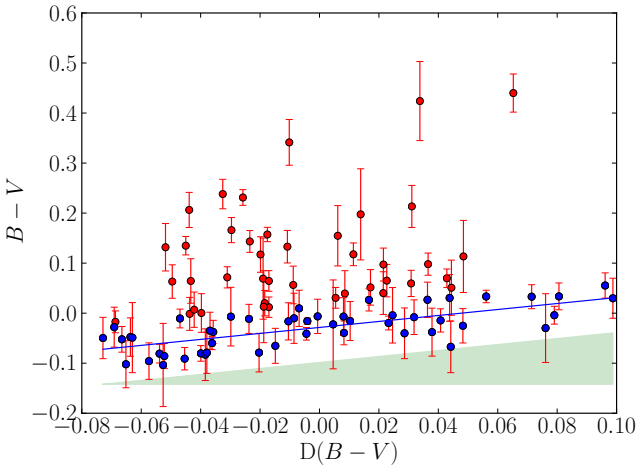

Figure 19. Same as Figure 17 for the $B-V$ color. The red points belong to significantly reddened supernovae (E(B-V)$\gtrsim$ 0.1), and the blue points represent almost unreddened ones.

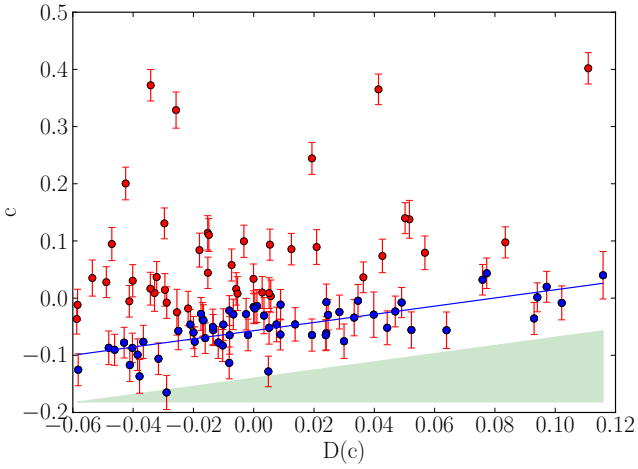

Figure 20. Same as Figure 19 for SALT2 color parameter c.

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 21 illustrates the correlation between Δm15(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 Δm15(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 22), for most of the objects. This is also reflected in the similar directions found to correlate with Δm15(B) and x_1 in Figure 12.

We emphasize that the correlations between directions in PC space and global photometric properties like x_1 and Δm15(B) represent yet another test of the information encompassed in the metric space. As it was constructed from the entire spectral sequences, it was expected to reproduce such photometric observables even though they were not inserted as features directly in the data matrix. This reinforces our statement that important information is preserved throughout the entire process.

6 CONCLUSIONS

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

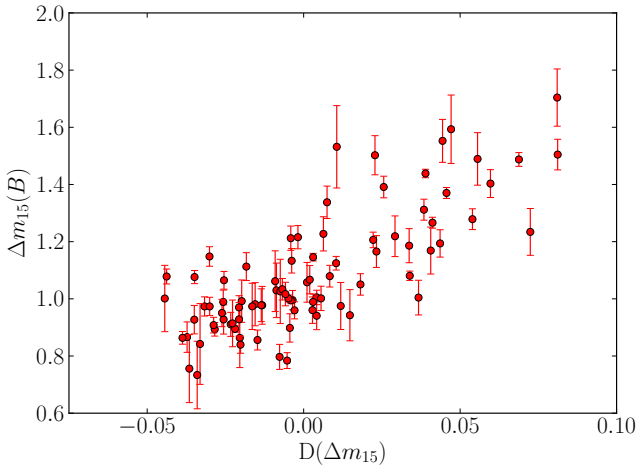

Figure 21. Correlation between PLS result and $\Delta m_{15}(B)$. The horizontal axis represents direction in 5D PC space which most correlates with $\Delta m_{15}(B)$ and the vertical axis is the value for this parameter measured from the SNe light-curves.

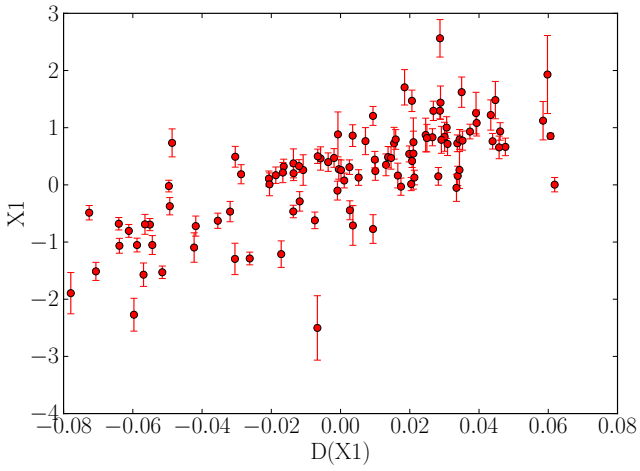

Figure 22. Same as Figure 21, but for the x_1 parameter of SALT2.

Square (PLS) techniques, it defines a meaningful metric space and correlates it to spectroscopic and photometric intrinsic properties.

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 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 use the available spectral evolution information.

We applied the method to a large sample ($\sim$ 120 SNe and $\sim$ 800 spectra) of well observed type Ia supernovae obtained by the SNfactory collaboration. At first, we defined a low dimensional parameter space using EMPCA and studied the spectral features covered by each PC separately. Results show that the High Velocity Features (HVFs) of Ca II H&K and infrared lines are uncorrelated with 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.

We confirmed many of the results of Cormier & Davis (2011). For example, the 91T-like SNe form a continuum of properties with normal SNe, PCA can be used to form a continuum of spectral templates, and the first two PCs mainly describe spectral velocities and equivalent widths. A larger dataset and the innovative method of analysing the derivative of the spectra allows 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 most used to sub-classify SNe Ia: velocity and pseudo-Equivalent Width (pEW) of Si II 5640 Å and Si II 6355 Å lines, B and V magnitudes and $B-V$ color at maximum, $\Delta m_{15}(B)$ and SALT2 parameters c and x_1. This demonstrates that the PC space is physically meaningful and includes the information recovered from usual spectral indicators. Moreover, it clarifies the potential of this framework in finding 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 by Foley & Kasen (2011). We showed that the velocity of the S II 5640 Å can be used for the same scope.

Now that we have the PCA trained on a large enough sample, this tool can be applied for direct comparison between synthetic and real SN Ia spectra. Projecting a synthetic spectral series in this PC space will reveal its counterparts among the real data, by an analysis of its neighbours. Moreover, the relations discovered by PLS can independently characterize each spectral feature of a given model and place it according to its most important physical properties within the real data parameter space.

Given the challenge of performing a coherent statistical comparison between synthetic and real spectra, our method can also be used to characterize complete sets of models built with 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 should be more robust against systematics in the models than comparing them with individual SNe. From a large enough synthetic spectra library,

the method also allows 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 applications will be investigated in a future work.

ACKNOWLEDGMENTS

We thank all the python, numpy and scipy communities for the high-quality free software they made available. MS thanks Philipp Edelmann for all the technical support during the developing of this work. EEOI acknowledges financial support from Brazilian agencies FAPESP (2011/09525-3) and CAPES (9229-13-2). Supported by German DFG Cluster of Excellence "Origin and Structure of the Universe" and the DFG Transregio Project 33 "Dark Universe". We thank Dan Birchall for observing assistance, the technical and scientific staffs of the Palomar Observatory, the High Performance Wireless Radio Network (HPWREN), and the University of Hawaii 2.2 m telescope. We recognize the significant cultural role of Mauna Kea within the indigenous Hawaiian community, and we appreciate the opportunity to conduct observations from this revered site. This work was supported by the Director, Office of Science, Office of High Energy Physics, of the U.S. Department of Energy under Contract No. DE- AC02-05CH11231; by a grant from the Gordon & Betty Moore Foundation; in France by support from CNRS/IN2P3, CNRS/INSU, PNCG, and the Lyon Institute of Origins under grant ANR-10-LABX-66 ; and in Germany by the DFG through TRR33 "The Dark Universe." Some results were obtained using resources and support from the National Energy Research Scientific Computing Center, supported by the Director, Office of Science, Office of Advanced Scientific Computing Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. HPWREN is funded by National Science Foundation Grant Number ANI-0087344, and the University of California, San Diego. This work was written using the collaborative ShareLaTeX platform.

APPENDIX A: 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 7) 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.

	$\sigma_{\rm res}$ training	$\sigma_{\rm res}$ 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 A1. Residuals in estimation of observables from training and validation samples.

Here we present results for $k = 10$ foldings, which is a standard first choice for many CV procedures (Arlot & Celisse 2009). 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 A1).

The stability of the PC space in $dF_{\log}$ is shown in Figures A1 and A2. The color code is the same used in Figure 5 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 A2, 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 A1.

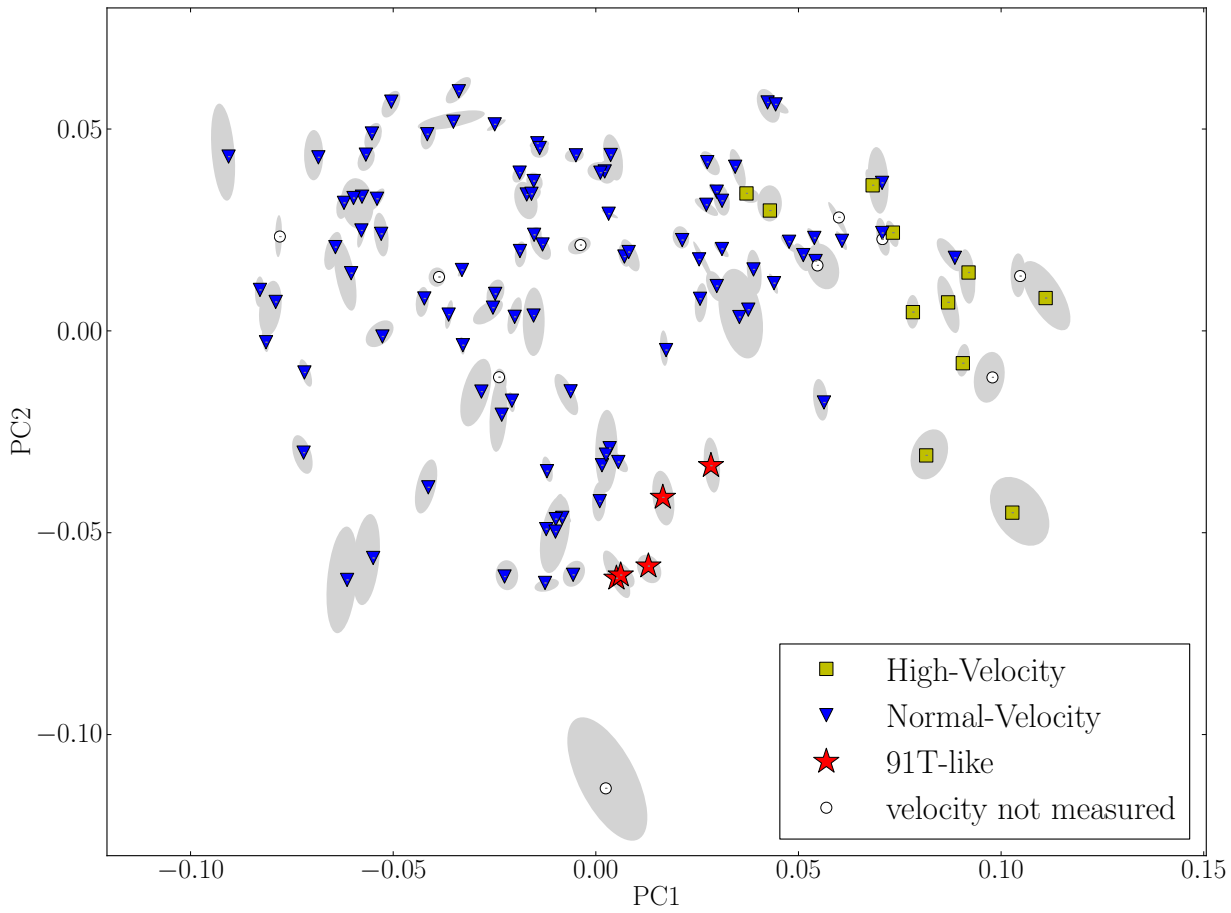

Figure A1. Stability of PC space through $k = 10$ folding cross-validation. The color code for points and line are the same used in Figure 5. The gray ellipses denote mean and 1σ variance for locations occupied by each data point throughout the 10 iterations.

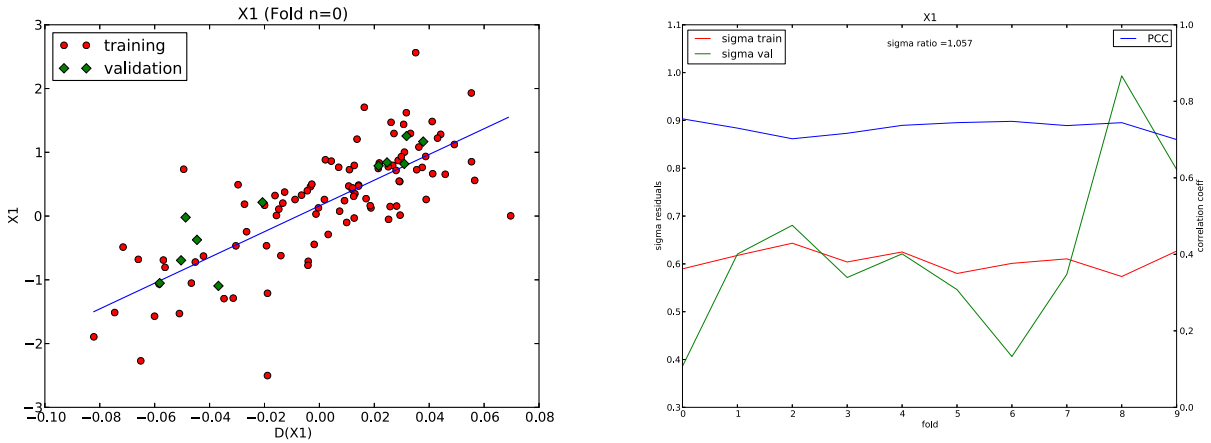

Figure A2. Accuracy of PLS analysis in predicting the value of x_1 for the validation sample. **Left panel:** Results for one of the realizations. The red circles and green diamonds correspond to the training and validation sets respectively. The blue line shows the result from the linear fit applied 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.

APPENDIX B: LINE VELOCITIES AND PSEUDO-EW 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. 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.

APPENDIX C: THE RECONSTRUCTIONS IN THE DERIVATIVE SPACE

In Figure C we show the same reconstructions presented in Figure 10 in the original derivative space. We lack a physical intuition in observing this space and it is hard to recognise 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 10 (SNF20071015-000 and SN2007kk), are not noticeable anymore.

REFERENCES

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., 2009, ArXiv e-prints

Arsenijevic V., Fabbro S., Mourão A. M., Rica da Silva A. J., 2008, A&A, 492, 535

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

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., 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

Buton C. et al., 2013, A&A, 549, A8

Cardelli J. A., Clayton G. C., Mathis J. S., 1989, ApJ, 345, 245

Childress M. et al., 2013, ApJ, 770, 107

Chotard N. et al., 2011, A&A, 529, L4

Connolly A. J., Szalay A. S., Bershady M. A., Kinney A. L., Calzetti D., 1995, AJ, 110, 1071

Copin Y. et al., 2006, NewAR, 50, 436

Cormier D., Davis T. M., 2011, MNRAS, 410, 2137

Dempster A. P., Laird N. M., Rubin D. B., 1977, JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B, 39, 1

Diemer B., Kessler R., Graziani C., Jordan, IV G. C., Lamb D. Q., Long M., van Rossum D. R., 2013, ApJ, 773, 119

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

Guy J. et al., 2007, A&A, 466, 11

Hachinger S., Mazzali P. A., Benetti S., 2006, MNRAS, 370, 299

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

Hügelmeyer 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

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

Jollife I. T., 2002, Principal Component Analysis. Springer-Verlag

Kromer M., Sim S. A., 2009, MNRAS, 398, 1809

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

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

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

Nugent P., Phillips M., Baron E., Branch D., Hauschildt P., 1995, ApJL, 455, L147

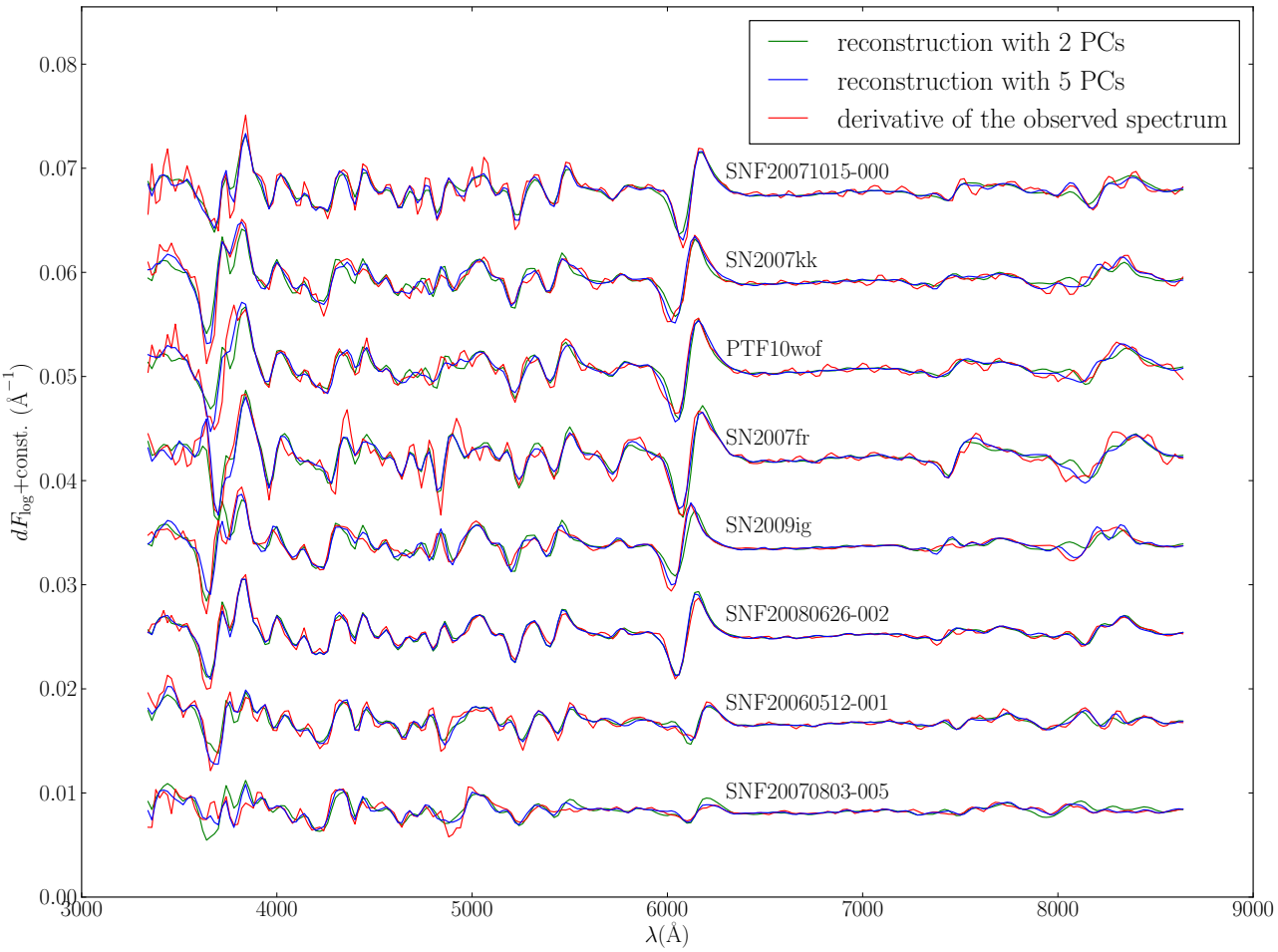

Figure C1. Comparison between the derivative of the observed spectra (red) and reconstructions from PCA using 2 (green) and 5(blue) PCs for a few supernovae at $B-$band maximum light.

Pakmor R., Kromer M., Taubenberger S., Sim S. A., Röpke F. K., Hillebrandt W., 2012, ApJL, 747, L10

Pedregosa F. et al., 2011, Journal of Machine Learning Research, 12, 2825

Pereira R. et al., 2013, A&A, 554, A27

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

Poznanski D., Nugent P. E., Filippenko A. V., 2010, ApJ, 721, 956

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. et al., 2012, ApJL, 750, L19

Roweis S., 1998, Advances in neural information processing systems, 626

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

Seitenzahl I. R. et al., 2013, MNRAS, 429, 1156

Silverman J. M. et al., 2012, MNRAS, 425, 1789

Sim S. A. et al., 2013, MNRAS, 436, 333

Stehle M., Mazzali P. A., Benetti S., Hillebrandt W., 2005, MNRAS, 360, 1231

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

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

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

Yip C. W. et al., 2004, AJ, 128, 2603