*Astroinformatics*
*Proceedings IAU Symposium No. 325, 2017*
*M. Brescia, G. Djorgovski, E. Feigelson,*
*G. Longo & S. Cavuoti, eds.*


DOI: 00.0000/X000000000000000X

# Exploring the spectroscopic diversity of type Ia supernovae with Deep Learning and Unsupervised Clustering

**Emille E. O. Ishida[1] and Michele Sasdelli[2] and Ricardo Vilalta[3] and Michel Aguena[4] and Vinicius C. Busti[4] and Hugo Camacho[4] and Arlindo M. M. Trindade[5,6] and Fabian Gieseke[7] and Rafael S. de Souza[8,9] and Yabebal T. Fantaye[10] and Paolo A. Mazzali[2]**

[1]Clermont Université, Université Blaise Pascal, CNRS/IN2P3, Laboratoire de Physique Corpusculaire, BP 10448, F-63000 - Clermont-Ferrand, France
email: emille.ishida@clermont.in2p3.fr

[2]Astrophysics Research Institute, Liverpool John Moores University, Liverpool L3 5RF, UK
[3]Department of Computer Science, University of Houston, 4800 Calhoun Rd., Houston TX 77204-3010, USA
[4]Departamento de Física Matemática, Instituto de Física, Universidade de São Paulo, CP 66318, CEP 05508-090, São Paulo - SP, Brazil
[5]Instituto de Astrofisica e Ciências do Espaço, Universidade do Porto, CAUP, Rua das Estrelas, PT4150-762 Porto, Portugal
[6]Departamento de Física e Astronomia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre 687, PT4169-007 - Porto, Portugal
[7]Institute for Computing and Information Sciences, Radboud University Nijmegen, Toernooiveld 212, 6525 EC - Nijmegen, Netherlands
[8]MTA Eötvös University, EIRSA "Lendulet" Astrophysics Research Group, Budapest 1117, Hungary
[9]IAG, Universidade de São Paulo, Rua do Matão 1226, 05508-900, São Paulo, Brazil
[10]Department of Mathematics, University of Rome Tor Vergata, Rome, Italy

**Abstract.** The existence of multiple subclasses of type Ia supernovae (SNeIa) has been the subject of great debate in the last decade. In this work, we show how machine learning tools facilitate identification of subtypes of SNe Ia. Using Deep Learning for dimensionality reduction, we were capable of performing such identification in a parameter space of significantly lower dimension than its principal component analysis counterpart. This is evidence that the progenitor system and the explosion mechanism can be described with a small number of initial physical parameters. All tools used here are publicly available in the Python package `DRACULA` (Dimensionality Reduction And Clustering for Unsupervised Learning in Astronomy) and can be found within COINtoolbox (`https://github.com/COINtoolbox/DRACULA`).

**Keywords.** supernovae: general – methods: machine learning, data analysis, statistical

## 1. Introduction

Type Ia supernovae (SNe Ia) are extremely bright objects, exhibiting a good degree of spectroscopic and photometric homogeneity. These properties allowed SNe Ia to play a major role in the discovery of the accelerating expansion of the Universe in the late $20^{\text{th}}$ century. However, despite a remarkable similarity among the bulk of SN Ia data, there is a significant fraction of spectroscopically peculiar objects. At this moment, it is still unclear if there exists different spectral subclasses of SN Ia, or if subtypes defined in the literature are just extremes of a continuum defined over the space of property distributions.

Parallel to such considerations derived from observational data, theoretical developments have investigated multiple hypotheses to explain SN Ia diversity. Proving the existence of well-defined and distinct subclasses would strongly support the different progenitor systems or, qualitatively, different scenarios for explosion mechanisms.

In this paper we use a series of machine learning tools and demonstrate how they can help automatize the visualization and classification of a large set of SN Ia spectra. Our method is composed of two steps: dimensionality reduction via deep learning, and clustering using K-means. Our goal is to provide a proof of concept, showing that the algorithm is able to leverage the same set of spectral features one would choose through visual recognition. This work was originally reported in Sasdelli *et al.* 2016 and the tools used here are implemented in the `DRACULA` Python package (Dimensionality Reduction And Clustering for Unsupervised Learning in Astronomy) and are publicly available within the COINtoolbox†.

## 2. Data

We compiled a set of publicly available SN Ia spectra from a variety of sources (see Sasdelli *et al.* 2016 for a complete reference list). These were smoothed with a Savitzky-Golay filter and derived as described in Sasdelli *et al.* 2015. Our data matrix set contains 3677 derivative spectra (lines) and 300 wavelength bins (columns) from which ≈150 spectra correspond to observations of different SNe Ia at maximum brightness.

## 3. Methodology

The first step in our algorithm is to reduce the initial dimensionality of the wavelength parameter space using Deep Learning (DL). In this context, the input data is made of feature vectors $\{\mathbf{x}\}$, where $\mathbf{x} = (x_1, x_2, ..., x_n)$. Each vector is represented as a layer of nodes, or neurons, where each node is a wavelength bin. Additional layers of neurons above the original input signal are built to ensure that each new layer captures a more abstract representation of the original input signal. This is achieved by computing at every node, a non-linear representation of the nodes in the layer below. The output is a new new vector $\mathbf{y} \in \mathcal{R}^d$, $d << n$, that abstracts the original representation, and is frequently better suited for classification. There are many approaches to implement a deep learning architecture; we follow the auto-encoder approach where the output vector is the result of encoding the original data through several layers of non-linear transformations (Vincent *et al.* 2008). It is important to emphasize that unlike the eigenvectors from principal component analysis, the elements of $\mathbf{y}$ do not follow a natural ordering.

Our goal is to characterize SNe Ia spectra at maximum brightness. However, as the number of available spectra at maximum is very small ($\approx$ 150 objects), performing dimensionality reduction in these data would not result in a robust low dimensional parameter space. On the other hand, we did have available a large number of spectra from other epochs (some of them have never been used before due to the lack of epoch determination). In order to take advantage of this large data set, we used *transfer learning* (TL), a recent area in machine learning that deals with the problem of leveraging information from a variety of different environments to help learn a model in a new context where training data is commonly scarce (Quinonero-Candela *et al.* 2009). In this paradigm, all available data is used in the construction of the low dimensional parameter space, but only those objects of particular interest are used in any subsequent analysis. Figure

† `https://github.com/COINtoolbox`

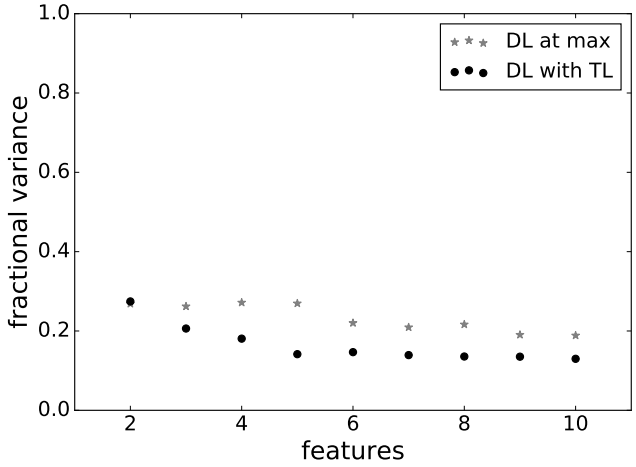


**Figure 1.** Fractional variance (difference between observed and reconstructed spectra) as a function of the dimensionality of DL space. Gray crosses show results from SNe Ia at maximum (no transfer learning) and black circles correspond to outcomes from all epochs (with transfer learning).

1 shows how using this approach improves the reconstruction power of the DL feature space. The plot shows the difference between reconstructed and observed spectra as a function of the number of DL features. A stable reconstruction power is achieved with only 4 DL dimensions, which is equivalent to 10 DL features when no transfer learning is invoked. In summary, we used all available SNe Ia spectra (3677) as an input for the DL algorithm, which provided a stable 4-dimensional parameter space. Afterwards, only the spectra at maximum (150) were used to investigate the existence of subgroups.

We used the K-Means clustering algorithm to test the feasibility of identifying subgroups of SN automatically. It is one of the most well-known clustering techniques (MacQueen 1967). The main idea is to find $k$ centres as cluster representatives. Starting with $k$ random centres, each data point is assigned to the cluster represented by its closest centre candidate (adopting an Euclidean distance). Once the first set of clusters is defined, the centre candidates are updated to the centroid defined by all the members in each cluster. The process is repeated until the centroids turn stable (do not change with further iterations). K-means requires to state the number of clusters, $k$, *a priori*. In our case, we wish to demonstrate that our approach is able to grasp the main spectral features underlying the classification proposed by Wang *et al.* 2009 (which is composed by subtypes normal, high-velocity, 91T-like and 91bg-like SNe) and so we investigate the K-Means output when requiring 1 to 4 groups.

Our results are reported in the 4-dimensional space obtained through DL, which is much smaller than the original 300 wavelength bins. However, this is still difficult to visualize. In order to facilitate the comparison between the results we found using K-Means and the results reported in the literature, we used *isomaps* (Tenenbaum *et al.* 2000) - a technique which belongs to a broader class of dimensionality reduction methods known as manifold learning. It can be seen as an extension of a *multi-dimensional scaling* able to capture non-linear manifold structures. Its goal is to find a low-dimensional embedding of the data, such that the distance between any pair of two points is preserved. It was recently applied to stellar spectral classification by Bu *et al.* 2014. Here we used isomap to provide a 2-dimensional visualization of the 4-dimensional DL feature space.

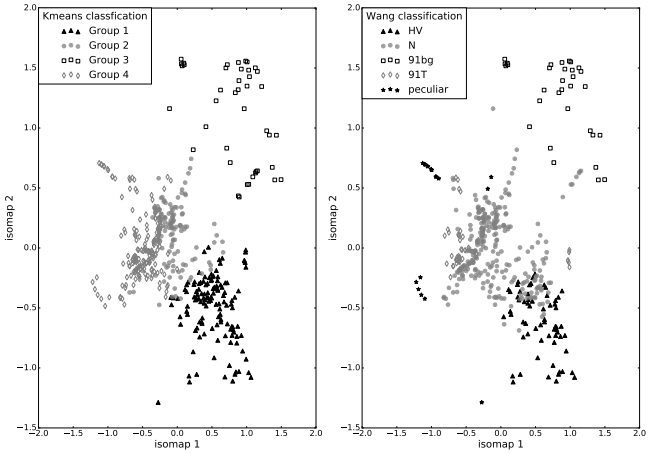


**Figure 2.** Four dimensional feature space from DL reduced to 2 dimensions through isomap. **Left:** Groups found by the K-Means algorithm when 4 groups are imposed. **Right:** Objects separated according to the classification proposed by Wang *et al.* 2009.

All of the algorithms described above, and a few more, are implemented in the `DRACULA` (Dimensionality Reduction And Clustering for Unsupervised Learning in Astronomy) Python package (Aguena *et al.* 2015). In the next section, we present the results obtained using this tool.

## 4. Results

Figure 2 shows the resulting 2-dimensional isomap space with the colours marking the clusters found by K-Means (left panel) and the SN Ia subtypes defined by Wang *et al.* 2009 (right panel). This figure not only demonstrates how isomaps can help high dimensional data visualization, but also clarifies the potential of combining DL with unsupervised learning algorithms. It also indicates that currently defined SNe Ia subtypes are extremes of a continuous distribution of spectral features.

In order to probe the ability of our method to identify the same data features found by visual inspection we go one step further and compare the spectral behaviour of each group in both configurations. Fig. 3 shows the mean spectrum of each cluster in both classification schemes. The agreement between them is proof that the combination of DL and K-Means is able to automatically identify important spectral features without human screening. The method recovers classes similar to those defined by visual inspection and can be used to optimize the current identification of subtypes of SNe Ia.

We also took advantage of this machinery to understand if there is an order among the physics underlying the definition of different SNe Ia subclasses. Depending on the degree of specialization we aim to achieve, it is possible to recognize a sequence of spectral characteristics which might be used to guide the physical basis of future data-driven classification systems. If we run the K-Means algorithm requiring only 2 groups, the resulting mean spectra resembles normal and high-velocity features SNe Ia. In case we require 3 groups, it is possible to recognize normal Ias, high-velocity and 1991bg-like spectra in the elements identified by our method. Finally, 1991T-like objects are separated from normals when 4 groups are required (Figure 4).

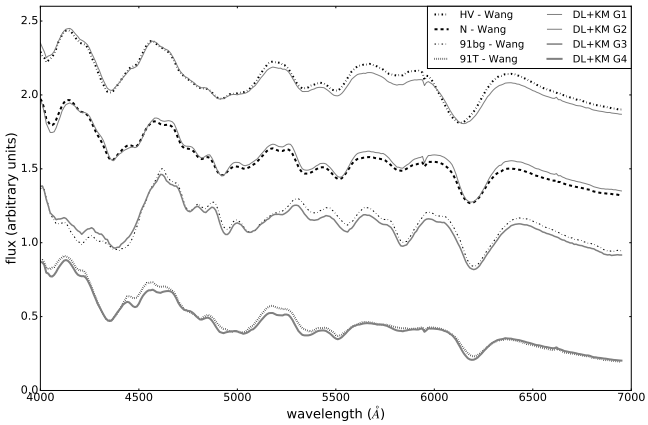


**Figure 3.** Mean spectrum found by DL + K-Means - full lines - and mean spectrum of the SNe Ia subtypes defined by Wang *et al.* 2009 - segmented lines.

## 5. Conclusions

We propose a framework to automatically identify subtypes of SNe Ia within a set of measured spectra by combining modern machine learning techniques. Our method can be summarized in 3 main steps: transfer learning, dimensionality reduction, and unsupervised learning.

The goal of transfer learning (TL) is to ensure the stability of the low dimensional feature space by adding a large variety of spectra in the original data matrix. In our particular case, we use all available SNe Ia spectra, irrespective of their epoch for determining a stable low dimensional parameter space and subsequently focused only on those spectra within $\pm 3$ days since maximum for the remaining analysis.

The dimensionality reduction stage was performed using Deep Learning (DL), which allowed us to translate an initial data matrix of 300 wavelength bins to a 4-dimensional feature space representation. Lastly, we use K-Means to investigate the possibility of identifying spectroscopic features in subclasses of SN Ia spectra at maximum light. This allows us to define a data-driven classification scheme and analyse clusters separately, looking for their spectroscopic characteristics. A more friendly visualization of the four dimensional feature space was achieved by the application of isomap as a further layer of dimensionality reduction.

We find that the spectral variability of SNe Ia at maximum can be summarized by a 4 dimensional space. This suggests that not much more than 4 underlying physical parameters are necessary to describe SN Ia explosions and their spectroscopic variability, including the most "peculiar" objects, such as 91bg-like and the 02cx-like SNe.

SNe Ia are well known as a uniform class of objects. Our results independently confirm that progenitors should be a "simple" system with no more than a handful of initial parameters. We also show that the currently identified SN Ia sub-types are extremes of a continuous distribution of spectral features and did not find strong evidence of distinct subclasses.

Our complete software apparatus was built under `DRACULA`, a publicly available Python package that is here tested on a public data set of SN Ia spectra. We refer the reader to Appendix B of Sasdelli *et al.* 2016, where a complete tutorial is presented. Although here we focus in the SNeÍa characterization problem, there is still great potential for application of this kind of tool in astronomical problems.

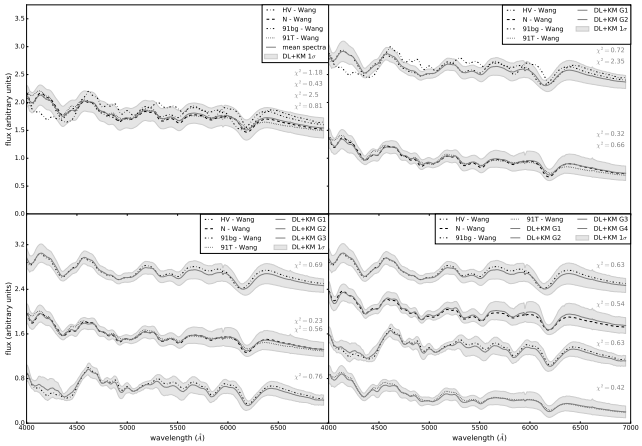


**Figure 4.** Sequence of groups found by K-Means (full lines) and the groups defined by Wang *et al.* 2009 (segmented lines). Panels run through 1 to 4 groups (top-left to bottom-right). We also report the deviations between each configuration mean spectra. The grey region denotes $1\sigma$ scatter for the groups defined with the K-Means algorithm.

Our results show agreement between mean cluster spectra and those proposed by Wang *et al.* 2009. Our method also confirms previous statements that high velocity features are of the first order effect in separating currently available samples of SNe Ia Wang *et al.* 2013. In subsequent work we will approach the time evolution in SN Ia spectra, and the possibility of using our tool in the classification of other supernova types.

## Acknowledgements

This work is a product of the 2$^{\text{nd}}$ COIN Residence Program. We thank Alan Heavens and Jason McEwen for encouraging the realization of this edition. The program was held in the Isle of Wight, UK in October/2015 and supported by the Imperial Centre for Inference and Cosmology (ICIC), Imperial College of London, UK, and by the Mullard Space Science Laboratory (MSSL) at the University College of London, UK. The IAA Cosmostatistics Initiative† (COIN) is a non-profit organization whose aim is to nourish the synergy between astrophysics, cosmology, statistics and machine learning communities. This work was written on the collaborative `Overleaf` platform‡.

## References

Aguena, M., Busti, V. C., *et al.* 2015, *Astrophysics Source Code Library* 1512.009
Bu, Y., Chen, F, *et al.* 2014, *New Astron.* 28, p. 35-43
MacQueen, J. 1967, in: *Proc. 5$^{th}$ Berkeley Symp. Math. Stat. and Prob.*, p. 281–297
Quinonero-Candela, J., Sugiyama, M., *et al.* 2009, in: *Dataset Shift in ML*, The MIT Press
Sasdelli, M., Hillebrandt, W., *et al.* 2015, *MNRAS* 447, Issue 2, p.1247-1266
Sasdelli, M., Ishida, E. E. O., *et al.* 2016, *MNRAS* 461, Issue 2, p. 2044-2059
Tenenbaum, J. B., Silva, V., *et al.* 2000, *Science*, 290, p. 2319-23
Vincent, P., Larochelle, H., *et al.* 2008, in: *Proc. 25$^{th}$ Intern. Conf. on ML*, Pages 1096-1103
Wang, X., Filippenko, A. V., *et al.* 2009, *ApJ* (Letters) 699, Issue 2, p. L139-L143
Wang, X., Wang, L., *et al.* 2013, *Science* 340, p. 170-173

† `https://asaip.psu.edu/organizations/iaa/iaa-working-group-of-cosmostatistics`
‡ `www.overleaf.com`