Statistical Challenges in 21st Century Cosmology
Proceedings IAU Symposium No. 306, 2014
A. F. Heavens, J.-L. Starck & A. Krone-Martins, eds.

doi:10.1017/S1743921314010928

Improved KPCA for supernova photometric classification

Emille E. O. Ishida[1], Filipe B. Abdalla[2] and Rafael S. de Souza[3,4]

[1]Max-Planck-Institut für Astrophysik,
Karl-Schwarzschild-Str. 1, D-85748 Garching, Germany
email: emille@mpa-garching.mpg.de

[2]Department of Physics and Astronomy, University College London,
London WC1E 6BT, UK
email: fba@star.ucl.ac.uk

[3]Korea Astronomy & Space Science Institute, Daejeon 305-348, Korea
[4]MTA Eötvös University, EIRSA "Lendulet" Astrophysics Research Group,
Budapest 1117, Hungary
email: rafael.2706@gmail.com

Abstract. The problem of supernova photometric identification is still an open issue faced by large photometric surveys. In a previous investigation, we showed how combining Kernel Principal Component Analysis and Nearest Neighbour algorithms enable us to photometrically classify supernovae with a high rate of success. In the present work, we demonstrate that the introduction of Gaussian Process Regression (GPR) in determining each light curve highly improves the efficiency and purity rates. We present detailed comparison with results from the literature, based on the same simulated data set. The method proved to be satisfactorily efficient, providing high purity ($\leqslant 96\%$) rates when compared with standard algorithms, without demanding any information on astrophysical properties of the local environment, host galaxy or redshift.

Keywords. (stars:) supernovae: general, techniques: photometric, methods: statistical

1. Introduction

In the late 20^{th} century, type Ia supernovae (SNe Ia) played a central role in the development of the standard cosmological model. Since then, a great effort has been employed towards the construction of a large, reliable SNe Ia sample, which might be able to provide further insights on the nature of the energy component driving the Universe accelerated expansion (dark energy). A major fraction of such efforts is directed to ongoing and planned photometric surveys, which will collect light curve information for thousands of SNe. However, the usefulness of these data to cosmology is limited by our ability to classify them without using spectroscopic information. As an example, the second SN data release from the *Sloan Digital Sky Survey* (Sako *et al.*, 2014) consists of 10258 variable and transient sources from which only 889 were spectroscopically confirmed.

In a previous work (Ishida & de Souza, 2013 - hereafter IdS2013), we demonstrated how the combination of Kernel Principal Component Analysis (kPCA) and Nearest Neighbour (NN) algorithms can be applied to the SN photometric classification problem, leading to competitive results specially when purity is the main feature to be maximise in the final photometric sample. In this work, we show how the introduction of Gaussian Process Regression (GPR) in the light curve fitting process can improve our previous results, allowing for better purity and efficiency rates.

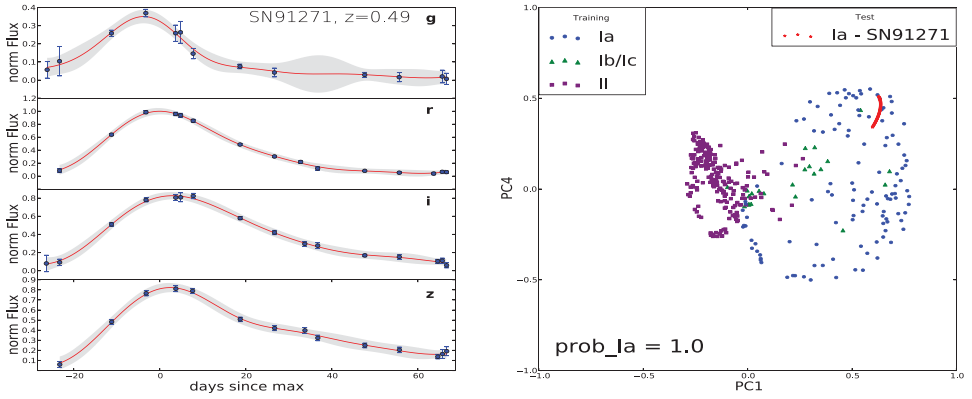

Figure 1. Results from the application of GPR to the pre-processing pipeline of supernovae data. **Left:** Normalized flux measurements as a function of days since maximum brightness. The blue dots denote observed values, the red line corresponds to the mean fitted light curve and the grey region encloses 2 standard deviations at each epoch. **Right:** Projection of supernovae light curves in PC space. The red path denotes the projection of 100 light curves defined within 2σ (grey region in the left panel).

2. Method

The framework presented in IdS2013 consists of applying kPCA to a light curve data set whose types are previously known (called *training set*) and subsequently use the geometrical distribution of SNe in this space to classify a new untyped SN by means of the NN algorithm. kPCA is a generalisation of PCA (see Jolliffe, 2012 for a complete review) built to handle the nonlinear case. It projects the data into a higher dimensional parameter space, where the relations can be linearly described, before determining the directions of each PC (for a deeper discussion see IdS2013 and references therein). After applying kPCA to the training set, the unclassified (or *test*) SN is projected into the PC space and it is assigned the same class as its first NN. As reported in IdS2013, this method achieves extremely high purity rates when applied to simulated data.

Nevertheless, it is important to emphasise that we need flux measurements equally sampled in time for all SNe in order to construct the initial data matrix. In IdS2013, we used a cubic spline interpolation in order to convert the observed flux measurements into a light curve function. Although this procedure proved to be enough to achieve competitive results, it does not enable us to propagate the uncertainty in flux measurements into the classification process. Here, we updated the method by introducing GPR as a light curve fitting technique.

GPR is a Bayesian approach to the linear regression problem, where we consider a prior distribution over the functions likely to describe the behaviour of our data. Before any measurements are taken, we consider a flat prior (e.g. for normalised light curves, $0 \leqslant \text{Flux} \leqslant 1$, $\forall$ epochs). Thus, any smooth function whose image is contained in this flux interval is as likely to describe our light curve. Once we have one measurement $y = F(\text{day}_1)$, it acts like a constraint over the previous prior and now we have a posterior probability distribution, enclosing all smooth functions within [0,1] interval, which pass through the observed point. Applying the same line of thought to successive measurements, we construct a region in the $y \times day$ parameter space, where all functions satisfying the posterior are concentrate (see Rasmussen & Williams, 2006). We used the GPR to determine a light curve function from the observed data points. This allowed

https://doi.org/10.1017/S1743921314010928 Published online by Cambridge University Press

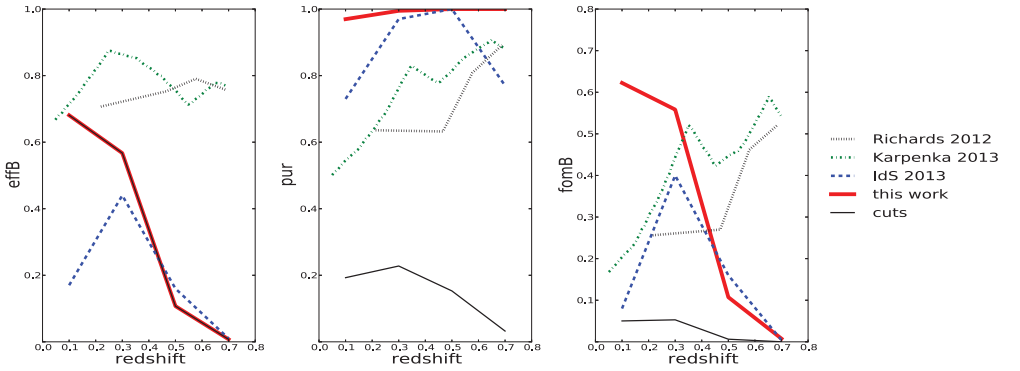

Figure 2. Comparison with literature results from different techniques, applied to the same data set, as a function of redshift. The black (dotted) line correspond to results from Richards *et al.*, 2012 (figure 10, $\mathcal{S}_{m,25}$), the green (dashed) line denotes those presented in Karpenka, Feroz & Hobson, 2013 (figure 3, D_4 without z) and the blue (bold-dashed) line shows our results when using a spline fit, reported in Ishida & de Souza, 2013 (table F2, D_7+SNR5). The red (bold-full) line represents results obtained with GPR, with data from -3 to +24 days, SNR$\geqslant$5 and 4-dimensional PC space. The thin line shows the outcomes when all SNe passing selection cuts are classified as Ia.

us to transfer the confidence region coming from the GPR directly into the classification process.

3. Application

We applied the procedure described above to the post-SNPCC data set†, which consists of a simulation corresponding to 5 years of observations from the *Dark Energy Survey*‡. From this initial set, we selected all objects with at least 3 observation epochs with signal to noise ratio, SNR$\geqslant$5 in each filter. Moreover, we demanded at least one observation before -3 and one observation after +24 days since maximum brightness. This reduced our sample to 465 and 3495 SNe in the training and test samples, respectively.

The left-hand panel of Figure 1 shows the results of applying GPR to a SN in the test sample, here we used a squared exponential covariance function and marginalized over the hyper-parameters. The right-hand panel illustrates the geometrical distribution of the training set in a 2-dimensional PC space¶ and the corresponding projections of 100 light curves fulfilling the constraints imposed by the posterior. We then applied the NN algorithm to all the 100 points individually, which enabled us to assign a probability of being a SN Ia to this SN. At a first glance, we notice the power of our method in separating Ia and non-Ia SNe. Although there is some degree of contamination, with the most drastic cases corresponding to type Ib/Ic SNe populating the area occupied by SN Ia.

Traditionally, results from such a classification analysis are reported in terms of efficiency (eff), purity (pur) and figure of merit (fom). Mathematically these are defined

† http://sdssdp62.fnal.gov/sdsssn/SIMGEN_PUBLIC/
‡ http://www.darkenergysurvey.org/

¶ Although the figure shows only 2 PC for didactic reasons, our final results were built in a 4-dimension parameter space. The choice of dimensionality and which PC to use was based on a cross-validation procedure as described in IdS2013, section 3.2.

https://doi.org/10.1017/S1743921314010928 Published online by Cambridge University Press

as

$$\text{eff} = \frac{N_{\rm cc}^{\rm Ia}}{N_{\rm tot}^{\rm Ia}}, \qquad \text{pur} = \frac{N_{\rm cc}^{\rm Ia}}{N_{\rm cc}^{\rm Ia} + N_{\rm wc}^{\rm nonIa}} \qquad \text{and} \qquad \text{fom} = \text{eff} \times \frac{N_{\rm cc}^{\rm Ia}}{N_{\rm tot}^{\rm Ia}(N_{\rm cc}^{\rm Ia} + W N_{\rm wc}^{\rm nonIa})}, \tag{3.1}$$

where $N_{\rm cc}^{\rm Ia}$, $N_{\rm tot}^{\rm Ia}$, $N_{\rm wr}^{\rm nonIa}$ are the number of correctly classified SNe Ia, the total number of SNe Ia in the test sample and the number of nonIa's which were wrongly classified as Ia, respectively. We assume a weight factor $W = 3$, which penalises wrong nonIa classifications contaminating the final sample (Kessler *et al.*, 2010). Following the nomenclature of IdS2013, the index "B" used after eff and fom indicates that $N_{\rm tot}^{\rm Ia}$ corresponds to the test sample *before* selection cuts. Thus, the effect of applied selection cuts are also taken into account.

Diagnostic results obtained from a GPR will always depend on the probability threshold (the minimum probability required so a SN is considered Ia). Given the statistical level of accuracy of current cosmological analysis, our goal is to maximise purity in our final sample, while keeping fom competitive. In this context, with a probability threshold of 0.95 and SNR$\geqslant$5, our method achieved 98% purity against 91% obtained with the cubic spline fit. The difference is still larger if we consider lower quality data. For SNR$\geqslant$3, GPR obtained purity of 97% while the cubic spline achieved no more than 67%. This proves that GPR is able to provide the same level of purity results even in the presence of lower quality data.

Figure 2 shows how our results compare to those of other methods available in the literature, when applied to the same data set. Here we show the best-case scenario pointed by the authors themselves. Richards *et al.*, 2012 used a combination of diffusion map and random forest algorithms and Karpenka, Feroz & Hobson, 2013 applied a neural network to the post-SNPCC data. Notice that GPR can improve efficiency and purity results over the spline fit even with a smaller light curve coverage (results shown from IdS2013 required data between [-3,+45] days since maximum brightness). Moreover, our method is the only able to achieve purity higher than 96% for the entire redshift range.

4. Conclusions

Our ability to extract cosmological information from future surveys is limit by how well we are able to photometrically classify the available SN samples. Given the current stage of cosmological analysis, where there is no consensus on how the purely photometric data should be used, we believe that our approach is ideal to select a small, high purity SN sample which might help improve the cosmological results.

Acknowledgements

EEOI is partially supported by Brazilian agency CAPES, grant number 9229-13-2. FBA acknowledges the Royal Society for a Royal Society University Research Fellowship.

References

Ishida, E. E. O. & de Souza, R. S. 2013, *MNRAS*, 430, 509
Jolliffe, I. T., 2002, Principal Component Analysis, *Springer-Verlag New York, Inc.*
Karpenka, N. V., Feroz, F., & Hobson, M. P. 2013, *MNRAS*, 429, 1278
Kessler, R., *et al.* 2010, *PASP*, 122, 1415
Rasmussen, C. E., & Williams, C. K. I. 2006, GP for Machine Learning, *MIT Press*
Richards, J. W. *et al.* 2012, *MNRAS*, 419, 1121
Sako, M. *et al.* 2014, `arXiv:`astro-ph/1401.3317