Procedia Computer Science

Volume 53, 2015, Pages 29–38

2015 INNS Conference on Big Data

# Discriminating Variable Star Candidates in Large Image Databases from the HiTS Survey Using NMF

Pablo Huijse[1,2], Pablo A. Estévez[2,1], Francisco Förster[3,1], and Emanuel Berrocal[1]

[1] Millennium Institute of Astrophysics, Santiago, Chile
[2] Department of Electrical Engineering, Universidad de Chile, Chile
[3] Center for Mathematical Modelling, Universidad de Chile, Avenida Blanco Encalada 2120 Piso 7, Santiago, Chile

**Abstract**

New instruments and technologies are allowing the acquisition of large amounts of data from astronomical surveys. Nowadays there is a pressing need for autonomous methods to discriminate the interesting astronomical objects in the vast sky. The High Cadence Transient Survey (HiTS) project is an astronomical survey that is trying to find a rare transient event that occurs during the first instants of a supernova. In this paper we propose an autonomous method to discriminate stellar variability from the HiTS database, that uses a feature extraction scheme based on Non-negative matrix factorization (NMF). Using NMF, dictionaries of image prototypes that represent the data in a compact way are obtained. The projections of the dataset into these dictionaries are fed into a random forest classifier. NMF is compared with other feature extraction schemes, on a subset of 500,000 transient candidates from the HiTS survey. With NMF a better class separability at feature level is obtained which enhances the classification accuracy significantly. Using the NMF features less than 4% of the true stellar transients are lost, at a manageable false positive rate of 0.1%.

*Keywords:* Data mining, Astronomy, Supernovae, Machine learning, Non-negative matrix factorization

## 1 Introduction

Nowadays astronomy is changing to a data-driven science [5]. The advances in observing and processing technologies have allowed the development of deep and extent sky surveys. One of the most emblematic examples of this is the Large Synoptic Survey Telescope (LSST; [11]), currently under construction in Chile. The LSST will start operating in 2022, covering the whole southern hemisphere sky. The LSST will generate a 150 Petabyte image database, 40 Petabytes worth of object catalogs, and 2 million transient alerts per night. The classification and characterization of astronomical phenomena on these large streams of data is a challenge

Selection and peer-review under responsibility of the Scientific Programme Committee of INNS-BigData2015


doi:10.1016/j.procs.2015.07.276

tackled by the new fields of astrostatistics and astroinformatics [1]. These fields combine techniques from statistics, computer science and engineering for the development of robust methods for big-data astronomical problems. It is expected that computational intelligence methods will be key additions in these new areas [7]. Astronomy is entering the era of big-data and preparations should be made in order to cope with the challenges imposed by surveys such as the LSST. In this paper we present a methodology to discriminate variable stars on large image catalogs from the High Cadence Transient Survey (HiTS) [6]. First the dimensionality of the dataset is reduced by projecting it to a set of prototypes obtained using Non-negative Matrix Factorization (NMF; [9, 3]). After that, the coefficients associated to the prototypes are fed to a classifier based on Random Forests [2]. The work presented in this paper considers a subset of 500,000 HiTS transient candidates found in 2013. We show that using the NMF features a better classification accuracy is obtained with respect to other feature extraction schemes. The dictionaries obtained by NMF are part-based and naturally sparse, which helps to improve class separability at feature level.

## 1.1 Astronomy Background

The HiTS (2013-2015), led by Francisco Förster *et al.*[1], is focused on the real time detection of supernovae (SNe), *i.e.* the explosion that characterizes the end of the life-cycle of a massive star [13]. The scientific objective of HiTS is to detect a supernova shock breakout (SBO), an event that occurs within hours after the supernova process begins. Detecting this event in the visible spectrum may help to prove or discard the theoretical models on SNe evolution (e.g. [14]). HiTS has been very successful on its SNe hunt [6], but the SBO is yet to be found. The HiTS observations were performed at Cerro Tololo, Chile using the Dark Energy Camera (DECam; [4]). DECam contains 60 CCD detectors of 2K x 4K pixels covering 3 square degrees of the sky. In optics the capabilities of a telescope are summarized by its "etendue", the product of the capture area in square meters and the camera field of view in square degrees. The etendue of DECam is 34 $m^2 deg^2$, currently the second largest in the world, and it corresponds to a 10% of the expected etendue of the LSST. HiTS observed 50 DECam fields (150 $deg^2$) with a cadence of 1.6 hours for 6 nights, between February and March 2015 [6]. The survey performed very deep observations in order to gather a large volume of events and increase the SBO detection chance. But the further the object the fainter it appears, making it harder to discriminate. Approximately a 66% of the interesting events are expected to be found between signal-to-noise ratio (SNR) 10 and 5 (the lower limit of the instrument).

The HiTS reduction pipeline performs real-time **image subtraction** in order to detect variable sources. For each field a given epoch is chosen as reference, to which all the other epochs are compared. In what follows the first steps of the HiTS pipeline are briefly described. In the first step an image is aligned to the reference by detecting point sources using standard astronomical tools and performing a second-order transformation. In the second step the point spread function (PSF) of the images are matched. When observing an astronomical object, what one actually gets is the convolution of the point source (object) with the PSF. The PSF takes into account effects of the instruments and the atmosphere that, in practice, broadens the object in pixel space. If the PSFs of the images are not matched before subtractions, artifacts will appear. In the third stage the subtraction is performed. After that, the difference image is divided by the local noise in order to obtain an "SNR image". Any region in the difference

[1] A manuscript titled "The High cadence Transient Survey (HiTS): real-time detection of supernovae and other transients using DECam" is in preparation. This manuscript will expose the HiTS tools, methods and results from 2013 to 2015.

image space, whose weighed integrated flux is larger than 5 standard deviations, is selected as a transient candidate. Each candidate is represented by a set of 21 x 21 pixel stamps obtained from the subtraction procedure and centered on the transient event. At this stage the candidates are dominated by statistical fluctuations, cosmic rays, and artifacts due to misalignments and wrong PSF matching. A training set using known variable stars artificially placed on DECam images was constructed. The database contains roughly a hundred million candidates.

### 1.2 NMF Background

Non-negative matrix factorization (NMF) [9] is a technique to learn localized representations for data in an unsupervised way. In NMF, a non-negative data matrix $V \in \mathbb{R}^{M \times N}$, containing N samples and M dimensions, is decomposed as $V \approx WH$, where $W \in \mathbb{R}^{M \times K}$ and $H \in \mathbb{R}^{K \times N}$ are the dictionary and coefficients matrices, respectively. The columns of the dictionary matrix represent the basis vectors of the decomposition and we will refer to them as **prototypes**. $K$ is a user defined parameter. The key difference between NMF and other decomposition techniques is the non-negativity constraint, which forces the components to be positive, but most importantly it forces the combinations between dictionary elements to be purely additive. This produces decompositions that are part-based and inherently sparse [9]. In this work we use Fast Hierarchical Alternating Least Squares (Fast HALS) [3], to solve the following constrained NMF optimization problem:

$$\min_{W,H} \|V - WH\|_F^2 + \lambda \|H\|_1 \text{ s.t. } W \geq 0, H \geq 0, \tag{1}$$

where a sparsity constraint for the coefficients based on the $\ell_1$ norm is used. The parameter $\lambda$ represents the trade-off between reconstruction error and sparsity. The Fast HALS algorithm performs a column-by-column optimization of the $W$ and $H^T$ matrices, reducing the computational time and memory requirements substantially in comparison with other NMF implementations [3].

## 2 Methods

### 2.1 Description of the subset of the HiTS survey

The HiTS subset that is used to train and test the proposed methodology consists of 500,000 objects selected as transient candidates by the HiTS survey in 2013. Each sample is represented by a 1323 dimensional vector corresponding to three 21 x 21 pixel images. The first image corresponds to the difference image divided by the local noise or **SNR image**. The transient candidate is centered in the SNR image. The second and third images correspond to the current and reference frame, from which the difference image was obtained. In what follows we refer to transient candidates associated to stellar variability as the true or positive class. Transient candidates associated to cosmic rays, detection and subtraction artifacts are referred as the false or negative class. The two classes in the 500,000 transient subset are balanced. Figure 1 show examples of positive and negative samples from the subset.

### 2.2 Description of the procedure to discriminate variable stars

In this section we describe the methodology used to discriminate variability due to stellar phenomena from the HiTS transient candidate database. First the transient subset is separated into a training set and testing set, each one having half the samples and preserving the balance

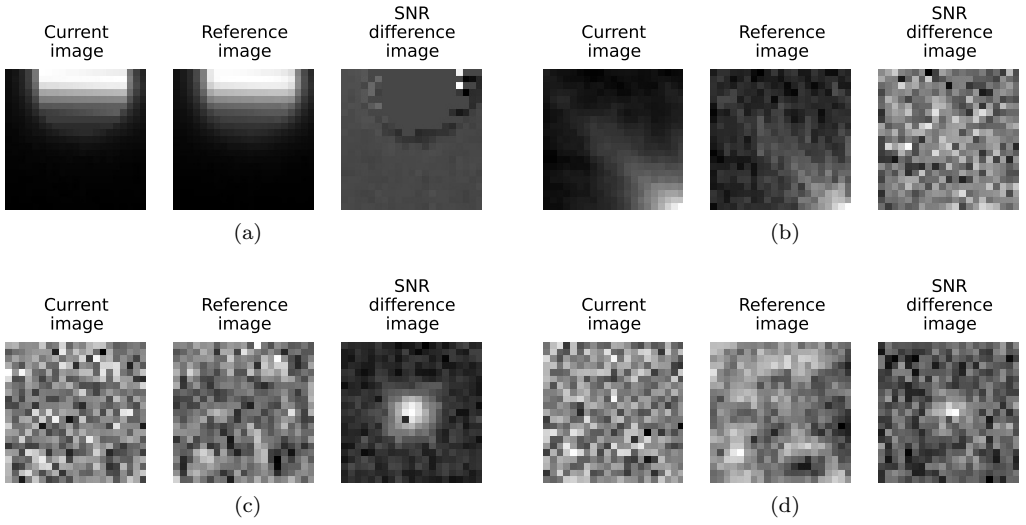


Figure 1: Examples from the 500,000 transient subset selected from the HiTS survey. For each sample (a-d), the current, reference and SNR difference image are showed. The colormap is scaled per image, so that the dimmest and brightest pixel correspond to start and end of the colormap, respectively. Figures (a) and (b) correspond to transients due to artifacts (false class). Figures (c) and (d) correspond to transients due to stellar variability (true class). The SNRs of the true samples are 50 and 6.

between classes. Ten disjoint partitions for the training and testing sets are used in order to obtain error bars. Three NMF dictionaries are built from the training set, one for each of the images contained in the samples. The HALS NMF procedure has two user-defined parameter, the number of prototypes in the dictionary ($K$) and the sparsity level of the decomposition ($\lambda$). The same number of prototypes is considered for the three dictionaries. HALS NMF is implemented in python[2]. Tests were performed on an Intel i5-4460 CPU with 16GB of RAM.

The method chosen to classify the samples is the random forest (RF) [2], primarily due to its competitive performance and simplicity in terms of parameter calibration. The RF implementation of the scikit-learn package is used [12]. The RF classifier is trained using an information gain criterion based on Shannon's entropy and 100 decision tree estimators. Because we are interested in testing NMF for feature extraction we do not delve much into parameter selection of the RF nor comparison with other classifiers. The procedure to obtain the dictionaries using NMF and train the classifier is as follows:

1. Select the number of prototypes ($K$) and the sparsity level of the decomposition ($\lambda$).
2. Normalize the images (SNR, current and reference) so that their pixels are mapped in $[0.0, 1.0]$. This is done for every single image and is required by the NMF algorithm to find meaningful structures.
3. Apply HALS NMF to the normalized training set, but doing it separately for each of the images comprising the samples. At the end of this step three dictionaries are gener-

[2]The python custom implementation of HALS NMF can be found online at `https://sites.google.com/site/pablohuijse/`.

ated, $W_1$, $W_2$ and $W_S$. Note that each column of the dictionary corresponds to a 21x21 prototype image.

4. Train a random forest classifier using the coefficient matrices of the three images, $H_1$, $H_2$ and $H_S$; plus the mean and standard deviation of each of the un-normalized images (*i.e.* having their original pixel values). This corresponds to a total of $3K + 6$ features.
5. Save the random forest model and the three dictionaries for further evaluation.

The dictionaries and the model are then evaluated on the test database as follows:

1. Normalize the samples so that each image has its pixels mapped in $[0.0, 1.0]$.
2. Project the new samples into the three dictionaries and obtain $\hat{H}_1$, $\hat{H}_2$ and $\hat{H}_S$. This is done using the HALS NMF algorithm but keeping the dictionaries fixed and update only the coefficients.
3. Obtain the mean and standard deviation of the original (un-normalized) SNR image, current image and reference image in the test database.
4. Set a probability threshold and evaluate the random forest model using the projected coefficients and statistical features.

For each trained classifier the predicted labels are compared to the real labels, to compute false positive rates (FPRs) and false negative rates (FNRs). We measure the performance of the method in terms of FPR versus FNR curves. Different discrimination threshold for the probability output of the classifier are used in order to obtain the curves. Results are evaluated in a high ($SNR > 6$) and low SNR ($5 < SNR \leq 6$) regimes.

# 3 Results

## 3.1 Parameter calibration

A 10-fold cross validation procedure is performed to select the optimal values of $K$ (number of prototypes) and $\lambda$ (sparsity). The following values for $K$ are considered: 10, 20, 50, 100 and 200. Note that $K$ prototypes are used per dictionary, so if $K = 200$ a total of 600 features is extracted. The following values for $\lambda$ are considered: 0, 5, 10, 25, 50, 100, 200. Our experiments show that the best classification performance is obtained with $K = 10$ and $\lambda = 0$, *i.e.* using no enforced sparsity. Figure 2 shows the performance curves for a selection of the tested $K$ and $\lambda$ values. We present this plot to illustrate two findings. First, we find that increasing $K$ reduces the performance of the classifier. This tell us that a few prototypes are needed to have a good characterization of the HiTS transient set. Also, by using fewer prototypes we are alleviating the "curse of dimensionality" which impacts the classification accuracy. Secondly, we find that the effect of the sparsity $\lambda$ is different depending on $K$. For lower values of $K$ the sparsity does not play an important role, but for larger values of $K$ including sparsity increases the performance. Remember that the sparsity on the coefficients (Eq. 1) represent a trade-off with the reconstruction error. When a small $K$ is used the sparsity may be redundant, as the images are already projected using few coefficients, and a better reconstruction might be preferable. On the other hand, when many prototypes are generated, the sparsity constraint yields more simple data reconstruction combinations, and the increased reconstruction error can be tolerated. Our observations hold for both SNR regimes. In what follows $K = 10$ and $\lambda = 0$ are selected as the best parameter combination for NMF.

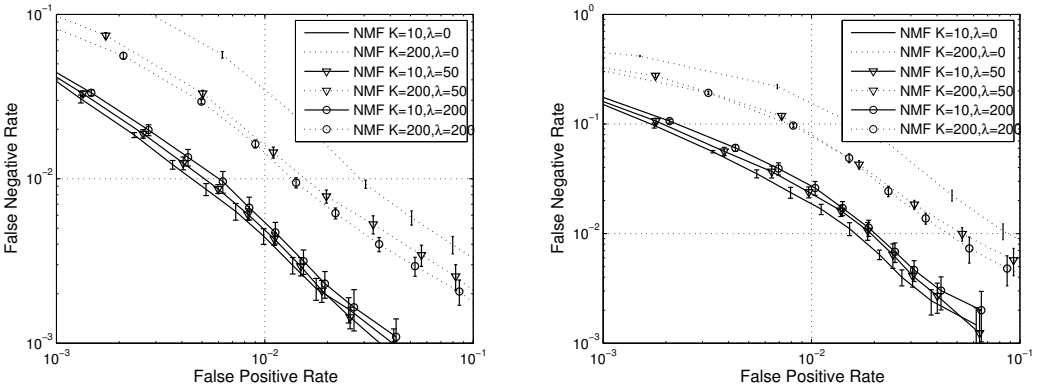


Figure 2: Curves of classification performance, using ten-fold cross validation, reducing dimensionality with NMF using different values of $K$ and $\lambda$, for $SNR > 6$ (a) and $SNR < 6$ (b). $K$ is to the number of prototypes per dictionary and $\lambda$ is the enforced sparsity.

## 3.2 Comparison with other feature extraction schemes

The performance of NMF is compared with Principal Component Analysis (PCA) and a scheme that uses the raw pixel data as input. PCA is a well-established method for dimensionality reduction. PCA finds an orthogonal transformation where the features (eigenvectors) are linearly uncorrelated. The new features can be sorted according to the amount of variance that it is explained from the data. Data dimensionality is reduced by reconstructing using only the most meaningful eigenvectors. We test the methodology proposed in Section 2.2 replacing NMF by PCA, and using different values of $K$. The best performance is obtained using $K = 10$, *i.e.*, preserving ten eigenvalues per dictionary, similarly to what was obtained with NMF.

Figure 3 shows a comparison between the best NMF model, the best PCA model and a classifier trained with the raw images as input (all-pixel model), *i.e.* no dimensionality reduction is performed in the latter case. The results show that the best classification performance is obtained by NMF, followed by PCA. This holds for both SNR regimes. If we consider a FPR of 0.1%, using NMF only a 4% of the true cases is miss-classified. On the other hand, for PCA and the all-pixel model, a 10.0% and a 15.6% is lost, respectively. For the faintest stars (low SNR), FNRs of 15.4%, 29.1% and 38.6% are obtained (at 0.1% FPR) for NMF, PCA and the all-pixel model, respectively. Note that in the final stage of the HiTS pipeline a team of astronomers manually inspect the classifier output for the positive class. False positives must be minimized, otherwise the manual inspection will be unmanageable. The trade-off of choosing a low FPR operation point, is that the faintest and most interesting events will be lost. The results show that the NMF model obtains the best results on both SNR scenarios, and incurs in the smallest classification performance penalty when the SNR decreases.

Figure 4 shows the 10 prototypes obtained by PCA (a) and NMF (b) in the SNR image dictionary. Figure 5 shows histograms of the most significant prototype for each sample. The first and second row correspond to PCA and NMF, respectively. The first and second column correspond to negative and positive samples, respectively. For PCA we can see that both distributions are very similar. Most of the samples are highly represented by the first and second prototype, which are the ones that explain the largest amount of variance in the model. For

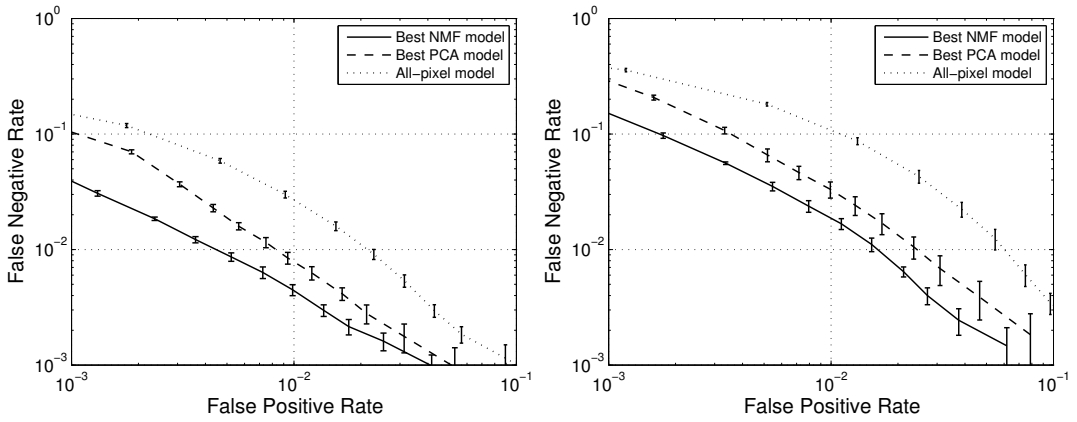


Figure 3: Classification performance in the test database of the proposed methodology based on NMF compared with similar implementations in which NMF is replaced by PCA (dashed line) and where no dimensionality reduction is performed, *i.e* all the pixels are used as features. Figures (a) and (b) correspond to high and low SNR regimes, respectively.

NMF, it can be observed that the distributions associated to the positive and negative samples are notably different. Most of the negative samples are represented by the first prototype, while the other prototypes follow an almost uniform distribution (except for the seventh and the ninth). Most of the positive samples are represented by the second, third, seventh and ninth. Even though, NMF is an unsupervised method, *i.e.* labels are not used, we observe that there is significant class separability in the features alone. This yields a better classification accuracy as shown before. The results observed for SNR image dictionary hold for the other dictionaries, but due to space limitations these results are omitted.

### 3.3 Dealing with large volumes of data

The HiTS transient candidate database contains roughly a hundred million samples. In order to apply the proposed method on the full HiTS database the issue of the scalability of the computations in terms of processing time and memory usage needs to be addressed. By using NMF we compress the data into low rank approximations, retaining only the most meaningful features. This improves the classification accuracy and also reduces the memory and time required to train and evaluate the classifier. In our case, using $K = 10$ we are reducing the required memory in almost two orders of magnitude. The Fast HALS method optimizes the NMF factors in a column-by-column way and employs by-products that depend on $K$ [3]. Compared to other NMF implementation Fast HALS is faster and less memory-bound. To further decrease computational time we compute matrix operations using optimized multi-core linear algebra libraries (OpenBLAS). It is also worth noting that projecting samples to an already trained dictionary has a low computational cost. Hence, testing subsets that are much larger than the training set can still be evaluated very efficiently. For very large databases one may train the NMF models with several relatively small but representative subsets of the database, using resampling techniques [8]. Distributed NMF implementations [10] that follow the MapReduce model are tailored for this kind of scenario and will be studied in the future.

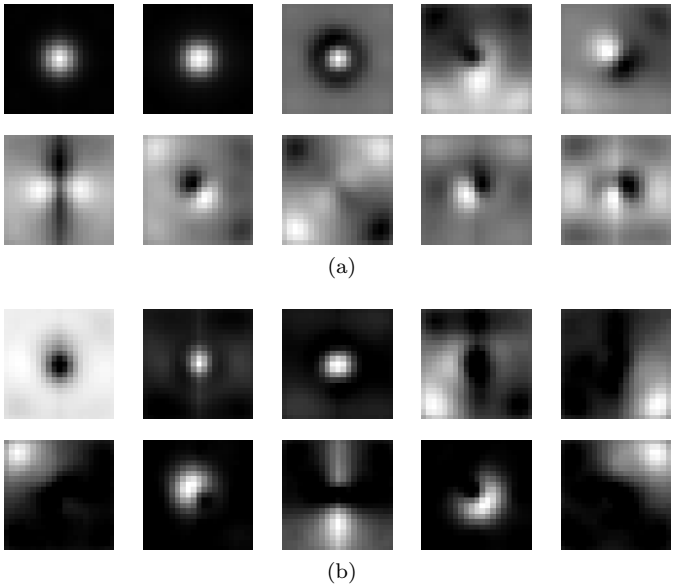

Figure 4: Ten prototypes obtained using PCA (a) and NMF (b). The PCA prototypes are more "holistic", *i.e.* they contain information of the whole image. The NMF prototypes represent parts of the images, such as structures near the borders, center and halos.

# 4 Conclusions

Astronomical surveys such as HiTS are generating several TeraBytes of image and catalog data. Within a few years, the LSST will be operational generating even more data and millions of transient alerts per night. Methods from the computational intelligence field are key to detect events on these large streams of data that otherwise might go unnoticed by human experts. The contribution of this paper is an autonomous and efficient methodology for stellar transient discrimination from images of the HiTS survey. This methodology relies on a feature extraction scheme based on NMF. The best classification performance is obtained using ten prototypes per dictionary, reducing the dimensionality from 1323 to 36. By inspecting the NMF prototypes we can appreciate that they represent parts of the images, which is due the non-negative constraints. This property is also responsible for the inherent class separability at feature level that is observed. For a given FPR, the NMF features achieve a lower FNR than PCA and the raw-pixel model. Particularly, when the NMF features are used, less than 4% of the true stellar transients are lost, at a manageable FPR of 0.1%. The NMF features are also more robust at lower SNR, which favors the detection of fainter transients. Once the dictionaries have been obtained, the computational cost of projecting new data is very low. Future work includes implementing an optimized HPC version of the codes to run on distributed environments in order to test the full extent of the HiTS survey, comparing with other feature extraction schemes, and testing additional supervised classifiers and deep learning based approaches.

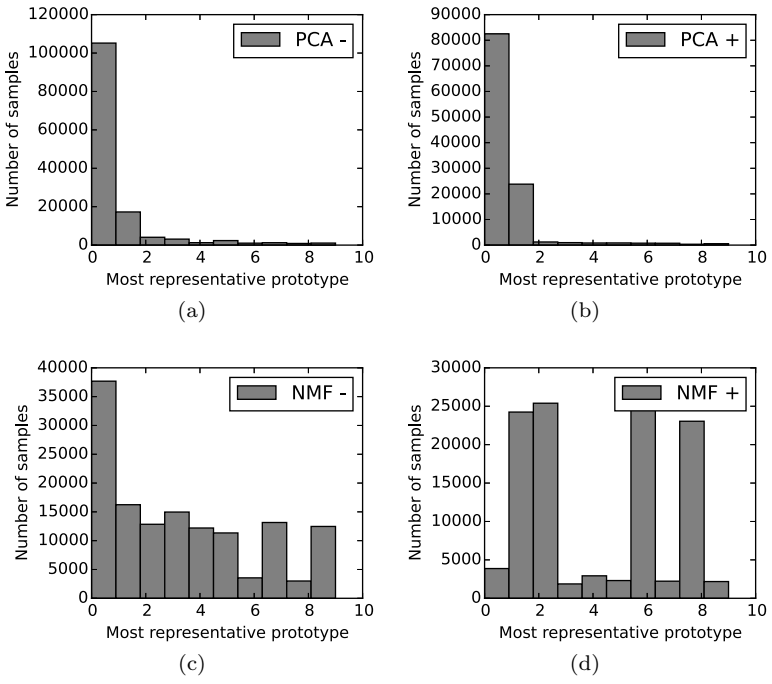


Figure 5: Histograms of the most significant prototype for each sample using PCA (a and b) and NMF (c and d). The first and second column correspond to the distribution of negative and positive samples, respectively. NMF is able to cluster the samples such that the classes are easier to separate. On the other hand the PCA distributions look very similar.

# 5 Acknowledgement

This work was funded by CONICYT-CHILE under grant FONDECYT postdoctoral $N^o$ 3150460, FONDECYT $N^o$ 11130228 and CONICYT NSF DPI 20140090. Pablo Huijse, Pablo Estévez and Francisco Förster acknowledge support from the Ministry of Economy, Development, and Tourism's Millennium Science Initiative through grant IC12009, awarded to The Millennium Institute of Astrophysics, MAS.

This project used data obtained with the Dark Energy Camera (DECam), which was constructed by the Dark Energy Survey (DES) collaborating institutions: Argonne National Lab, University of California Santa Cruz, University of Cambridge, Centro de Investigaciones Energeticas, Medioambientales y Tecnologicas-Madrid, University of Chicago, University College London, DES-Brazil consortium, University of Edinburgh, ETH-Zurich, Fermi National Accelerator Laboratory, University of Illinois at Urbana-Champaign, Institut de Ciencies de l'Espai, Institut de Fisica d'Altes Energies, Lawrence Berkeley National Lab, Ludwig-Maximilians Universitat, University of Michigan, National Optical Astronomy Observatory, University of Nottingham, Ohio State University, University of Pennsylvania, University of Portsmouth, SLAC National Lab, Stanford University, University of Sussex, and Texas A&M University. Funding for DES, including DECam, has been provided by the U.S. Department of Energy, National Science Foundation, Ministry of Education and Science (Spain), Science and Technology Facilities Council (UK), Higher Education Funding Council (England), National Center for Supercomput-

ing Applications, Kavli Institute for Cosmological Physics, Financiadora de Estudos e Projetos, Fundação Carlos Chagas Filho de Amparo a Pesquisa, Conselho Nacional de Desenvolvimento Científico e Tecnológico and the Ministério da Cincia e Tecnologia (Brazil), the German Research Foundation-sponsored cluster of excellence "Origin and Structure of the Universe" and the DES collaborating institutions.

# References

[1] K.D. Borne. Astroinformatics: Data-oriented astronomy research and education. *Journal of Earth Science Informatics*, 3:5–17, 2010.

[2] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[3] Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shun-ichi Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

[4] Thomas Diehl. The dark energy survey camera (DECam). *Physics Procedia*, 37(0):1332 – 1340, 2012. Proceedings of the 2nd International Conference on Technology and Instrumentation in Particle Physics (TIPP 2011).

[5] E.D. Feigelson and G.J. Babu. Big data in astronomy. *Significance*, 9(4):22–25, 2012.

[6] F. Forster et al. HiTS real-time supernova detections. *The Astronomer's Telegram*, 7099:1, February 2015.

[7] P. Huijse, P.A. Estevez, P. Protopapas, J.C. Principe, and P. Zegers. Computational intelligence challenges and applications on large-scale astronomical time series databases. *Computational Intelligence Magazine, IEEE*, 9(3):27–39, Aug 2014.

[8] Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael I Jordan. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816, 2014.

[9] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[10] Chao Liu, Hung-chih Yang, Jinliang Fan, Li-Wei He, and Yi-Min Wang. Distributed nonnegative matrix factorization for web-scale dyadic data analysis on mapreduce. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 681–690, New York, NY, USA, 2010. ACM.

[11] LSST Science Collaborations and LSST Project 2009. LSST Science Book, 2013. Version 2.0, arXiv:0912.0201, `http://www.lst.org/lsst/scibook`.

[12] F. Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[13] J.R. Percy. *Understanding Variable Stars*. Cambridge University Press, 2007.

[14] N. Tominaga, T. Morokuma, S. I. Blinnikov, P. Baklanov, E. I. Sorokina, and K. Nomoto. Shock Breakout in Type II Plateau Supernovae: Prospects for High-Redshift Supernova Surveys. *The astrophysical journal, supplement series*, 193:20, March 2011.