SF2A 2021
A. Siebert, K. Baillié, E. Lagadec, N. Lagarde, J. Malzac, J.-B. Marquette, M. N'Diaye, J. Richard, O. Venot (eds)

UNSUPERVISED CLASSIFICATION OF CIGALE GALAXY SPECTRA

J. Dubois[1], D. Fraix-Burnet[1], J. Moultaka[2], D. Burgarella[3] and P. Sharma[3]

Abstract. The goal of the work presented is to gain a thorough understanding of the unsupervised classification of galaxies' spectra using the Fisher-EM discriminative latent subspace Gaussian mixture technique to assess its physical relevance as well as to characterise the effect of the noise on the process. We simulated a sample of 11 475 optical spectra of galaxies with 496 monochromatic fluxes using the CIGALE tool. The integrated completed likelihood (ICL) criterion in Fisher-EM is used to determine the statistical model and the optimal number of clusters. The analysis was run numerous times to ensure that the results were reliable. The optimal classification obtained contains 12 clusters and is very robust against noise down to a signal-to-noise ratio (SNR) of 3. The distribution of the parameters used for the simulation shows excellent discrimination between classes. This study yields two conclusions valid at least for the Fisher-EM algorithm. Firstly, the unsupervised classification of spectra of galaxies is both reliable and robust to noise. Secondly, such analyses can extract the useful physical information contained in the spectra and build highly meaningful classifications. In an epoch of data-driven astrophysics, it is important to trust unsupervised machine learning approaches that do not require unavoidably biased training samples.

Keywords: Methods: data analysis – Methods: statistical – Galaxies: statistics – Galaxies: general – Techniques: spectroscopic

1 Introduction

Whether it be classifications based on morphological properties of galaxies, their spectra, or spectral features, the idea of classifying galaxies is not new and several studies were previously published on the matter (Kennicutt (1992), Dobos et al. (2012), Wang et al. (2018), Siudek et al. (2018)). Here we present a new, fully data-driven, and machine-learning-based approach for classifying galaxy spectra using the unsupervised classification method Fisher-EM (Bouveyron & Brunet (2011)). In this work, we have applied Fisher-EM on a large sample of optical galaxy spectra simulated using the spectral energy distribution (SED) fitting code CIGALE (Burgarella et al. (2005), Noll et al. (2009), Boquien et al. (2019)). We were able to assess the physical relevance of the method and the effect of noise on the process. We here provide a highlight of the most important results. In 2.1, we present the discriminative ability of the method in regards to the galaxies' physical properties. In 2.2, we show the robustness of the method against the presence of noise. A paper with a thorough description of the data, method, and results was submitted to A&A, and is undergoing the reviewing process at the time of writing.

2 Highlights of the classification

2.1 Discrimination of the galaxies' physical properties

The classification was performed on the sole basis of the spectra. The optimal classification contains 12 classes and is shown in Fig. 1 (left and center panel). As seen in the left panel, the dispersion of the spectra in the classes is relatively small, showing good homogeneity of the classes. Perhaps the most interesting result lies in the center panel, which shows that some of the galaxies' physical parameters in the classes (see Boquien et al. (2019) for more information about those parameters) are precisely discriminated. A linear discriminant analysis (LDA) shows that the most segregated parameters are the age of the galaxy T_{main}, the metallicity, the fraction of stars issued from a late burst of star formation f_{burst}, and the age of the late burst T_{burst}.

[1] Institut de Planétologie et d'Astrophysique de Grenoble, 38400 Saint-Martin d'Hères, France. email: julien.dubois@univ-grenoble-alpes.fr
[2] Institut de Recherche en Astrophysique et Planétologie, 31028 Toulouse, France
[3] Laboratoire d'Astrophysique de Marseille, 13013 Marseille, France

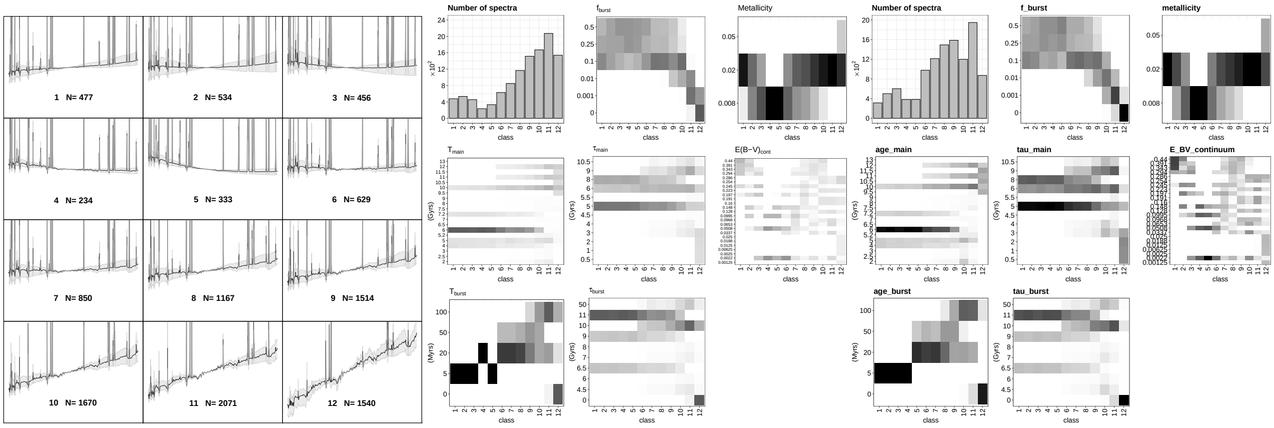

Fig. 1. Left: 12-clusters classification of the noiseless spectra. Mean spectra (in black) and their dispersion (in grey) for every class of the 12-clusters classification. The dispersion corresponds to the 10% and 90% quantiles for each monochromatic flux. **Center:** 12-clusters classification of the noiseless spectra. *Top left:* number of spectra contained in each class. *All others:* heatmaps of the relevant CIGALE input parameters among the 12 classes on noiseless spectra. All possible parameter values are represented on the y-axis, and the class index on the x-axis. The in-class densities of the parameter values are illustrated in the form of a heatmap, where a dark square equates to a density of 1, and white of 0. The classes are sorted by ascending average T_{main}. **Right:** Same but for spectra with an added noise of SNR=20.

2.2 Robustness against noise

The analysis was done multiple times with different levels of noise, characterised by their signal-to-noise ratio (SNR). The method showed consistent results for SNRs as low as 3. An example for SNR=20 is shown in the right panel of Fig. 1. At SNR=1, the optimal number of clusters drastically decreases down to 5 due to the loss of information, but the method is still capable of providing a certain degree of discrimination.

3 Conclusions

The study shows that the unsupervised classification algorithm Fisher-EM applied on thousands of CIGALE galaxy spectra yields a classification that is both robust against the initialisation of the algorithm and the noise. Very importantly, the classification is very discriminating with respect to the physical properties of the galaxies. Our findings provide considerable motivation to investigate the atlas presented in Fraix-Burnet et al. (2021) in greater depth and to apply it to larger samples. The interesting prospect is to incorporate galaxies at higher redshifts to investigate the classification's evolution over time using a fully data-driven procedure.

References

Boquien, M., Burgarella, D., Roehlly, Y., et al. 2019, A&A, 622, A103

Bouveyron, C. & Brunet, C. 2011, arXiv e-prints, arXiv:1101.2374

Burgarella, D., Buat, V., & Iglesias-Páramo, J. 2005, MNRAS, 360, 1413

Dobos, L., Csabai, I., Yip, C.-W., et al. 2012, MNRAS, 420, 1217

Fraix-Burnet, D., Bouveyron, C., & Moultaka, J. 2021, Astronomy and Astrophysics, in press

Kennicutt, Robert C., J. 1992, ApJS, 79, 255

Noll, S., Burgarella, D., Giovannoli, E., et al. 2009, A&A, 507, 1793

Siudek, M., Małek, K., Pollo, A., et al. 2018, arXiv e-prints, arXiv:1805.09905

Wang, L.-L., Luo, A. L., Shen, S.-Y., et al. 2018, MNRAS, 474, 1873