Mem. S.A.It. Vol. 94, 124

Memorie della

Clustering of galaxy spectra: an unsupervised approach with Fisher-EM

J. Dubois[1], D. Fraix-Burnet[1], and J. Moultaka[2]

[1] Univ. Grenoble Alpes, CNRS, IPAG, 38000 Grenoble, France e-mail: julien.dubois@univ.grenoble-alpes.fr
[2] IRAP, Université de Toulouse, CNRS, CNES, UPS, 14 avenue Edouard Belin, 31400, Toulouse, France

Received: 30-11-2022; Accepted: 21-02-2023

Abstract. We present a novel approach to galaxy spectra classification using Fisher-EM, a latent subspace clustering and Gaussian mixture model based algorithm. This approach was applied to a sample of 10 000 simulated spectra, highlighting its capacity to discriminate physical properties based on spectroscopic data, as well as its robustness towards noise. A sample of 700 000 spectra of close-by galaxies observed by the Sloan Digital Sky Survey (SDSS) was successfully classified, and a detailed physical interpretation of the classes is in preparation. An extension to higher redshifts is currently in progress, using a sample of 70 000 galaxies of redshift $0.4 < z < 1.2$ from the VIMOS Public Extragalactic Survey (VIPERS). An evolution tree-like structure was constructed, showcasing evolution pathways of the classes throughout cosmic time from $z = 1.2$ to $z = 0.4$.

Key words. Methods: data analysis – Methods: statistical – Galaxies: statistics – Galaxies: general – Techniques: spectroscopic

1. Introduction

The JWST is now in operation, and the world of astrophysics is about to enter a new era. Observations are made at greater redshifts, and now is a particularly exciting time to study galaxies and their evolution.

The motivation behind this work is to highlight the diversity and evolution of the physical aspects of galaxies that can be inferred from their spectroscopic properties (e.g. metallicity, history of star formation, active nuclei, etc.). In particular, working with a broad range of redshifts makes it possible to evaluate the evolution of these characteristics through the history of the universe, which is ultimately the goal of this work. Using an unsupervised clustering algorithm, an automatic classification of large samples of galaxy spectra is made possible (Fraix-Burnet et al. 2015). Galaxies sharing similar properties are gathered in classes, and the physical interpretation can thus be restricted to the classes' mean spectrum instead of the whole sample, hence greatly reducing the computational cost.

The work presented in this contribution consists of an overview of several applications of the unsupervised classification algorithm Fisher-EM to three different datasets of optical spectra. First, on a simulated sample (Dubois et al. 2022), then on observations from the Sloan Digital Sky Survey (SDSS) (Fraix-

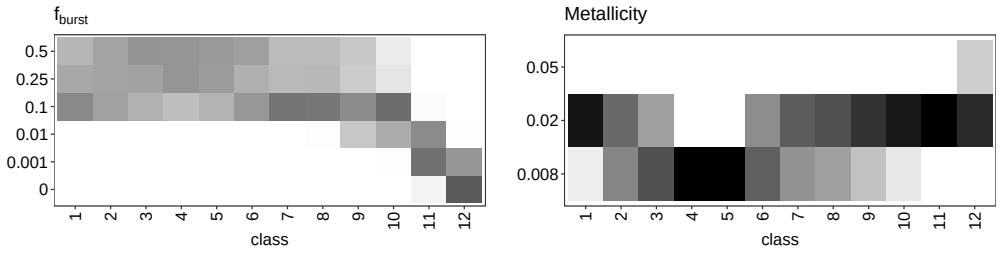

Fig. 1. Heatmaps showing the distribution of the mass fraction of stars originating from a sudden burst of star formation (left panel), and the metallicity (right panel) in the classification of a simulated sample of galaxy spectra. The parameter values are represented on the y-axis, and the class index on the x-axis. The within-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.

Burnet et al. 2021), and finally, on observations from the VIMOS Public Extragalactic Redshift Survey (VIPERS). Another contribution closely related to this work was made by M. Siudek. It showcases Fisher-EM applications to photometric data from VIPERS (Siudek et al. 2017, 2018).

The unsupervised algorithm is briefly explained in Sect. 2. An overview of the results that have been obtained is presented in Sect. 3 and Sect. 4.

2. Method

The algorithm Fisher-EM was used for this project. It combines both an unsupervised approach based on a Gaussian Mixture Model (GMM), and a dimension reduction process through projection of the data onto a discriminative latent subspace. The description of Fisher-EM is kept concise here, but a complete and thorough explanation can be found in the original article (Bouveyron & Brunet 2012).

A GMM models the data's probability density function (PDF) $f(K, \theta)$ with the weighted sum of K multivariate Gaussian PDFs $\Phi(\mu_i, \Sigma_i)$ (Eq. 1). The algorithm aims at minimizing a loss-function by tweaking the number of classes K, the shapes of the GMM's multivariate Gaussian PDFs $\Phi(\mu_i, \Sigma_i)$ and their weight π_i.

$$f(K, \theta) = \sum_{i=1}^{K} \pi_i \Phi(\mu_i, \Sigma_i) \tag{1}$$

An important aspect of this method to grasp is that the GMM is applied on a latent subspace rather than the observed space. It serves two purposes. It reduces the dimension down to $K - 1$; dimension reduction is necessary to avoid the curse of dimensionality that high-dimensional data suffer from. And in addition, the subspace is chosen to facilitate the convergence of the GMM. In fact, the projection matrix is tweaked iteration after iteration at the same time as the GMM parameters such that the Fisher criterion is maximized. This criterion is the ratio of the between-class variance over the within-class variance. As such, maximizing it amounts to maximizing the between-class variance (i.e. having the clusters well separated) and minimizing the within-class variance (i.e. having the clusters as compact as possible). This way, the subspace is optimized to highlight the discriminative features in the data and to make the clusters stand out.

3. Physical relevance and robustness of the classification

The physics of galaxies is encoded in their spectra; the presence of emission lines can for example be linked to episodes of star formation and activity in the galactic nuclei; the shape of the continuum gives an indication of the age of the stars; lots of different aspects in a spectrum can be linked back to many physical characteristics. However, a spectrum also contains a lot of non-physical particularities that

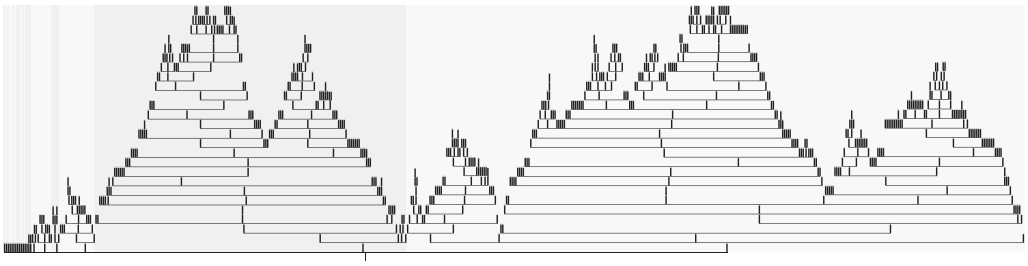

Fig. 2. This tree-like structure highlights evolution pathways of galaxy classes over cosmic time up to a redshift of $z = 1.2$. Each vertical step in the tree corresponds to a certain epoch, linearly sampled from 4 Gyr after the Big Bang (bottom of the tree) to 9 Gyr (top of the tree).

can influence the classification. For example, the amount of noise is not necessarily identical in every observation; there can be differences in calibration, sampling, and other instrumental aspects. And because all these components are mixed together in a spectrum, there was a need to test the ability of Fisher-EM to produce classifications that are physically relevant, i.e. that segregate galaxies based on physical characteristics.

Results on optical spectra simulated with CIGALE (Boquien et al. 2019) show that it is in fact the case (Dubois et al. 2022). Some physical characteristics are very well separated among the classes. The presence and the mass fraction of a burst of star formation, for example, is very well segregated (Fig. 1).

4. A window to galaxy evolution

The method was successfully applied to a sample of 700 000 spectra of close-by galaxies from the SDSS (Fraix-Burnet et al. 2021). This is the first automatic classification of a large sample of galaxy spectra performed without any prior feature selection. The resulting classification contains 86 classes, each showing their own specificities, and an in-depth description of the classes is currently in preparation.

An extension of the SDSS classification is currently in progress, and it shall include galaxies from the VIPERS PDR-2 up to a redshift of $z = 1.2$. The sample was divided into subsamples by bins of epoch, and each subsample was individually classified with Fisher-EM. Links between classes of successive epochs were constructed with k-NN to make it possible to follow the evolution of a given class over the cosmic time from 4 Gyr to 9 Gyr after the Big Bang.

When put together, the links create a tree-like structure, where branches can be followed to track evolution pathways (Fig. 2). Preliminary analysis shows a dozen of significant branches with clear spectral specificities, mostly divided into two main categories: star forming and passive. Morphological, spectral, and photometric measurements are being cross-matched with the classes, and will give precise insights for the interpretation of the evolution branches.

Acknowledgements. This paper uses data from the VIMOS Public Extragalactic Redshift Survey (VIPERS). VIPERS has been performed using the ESO Very Large Telescope, under the "Large Programme" 182.A-0886. The participating institutions and funding agencies are listed at http://vipers.inaf.it

References

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

Bouveyron, C. & Brunet, C. 2012, Statist. Comput., 22, 301

Dubois, J., Fraix-Burnet, D., Moultaka, J., Sharma, P., & Burgarella, D. 2022, Astronomy & Astrophysics, 663, A21

Fraix-Burnet, D., Bouveyron, C., & Moultaka, J. 2021, Astronomy and Astrophysics, 649, A53

Fraix-Burnet, D., Thuillard, M., & Chattopadhyay, A. K. 2015, Frontiers in Astronomy and Space Sciences, 2

Siudek, M., Małek, K., Pollo, A., et al. 2018, Astronomy & Astrophysics, 617, A70

Siudek, M., Małek, K., Scodeggio, M., et al. 2017, Astronomy & Astrophysics, 597, A107