Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **472,** 2808–2822 (2017)
Advance Access publication 2017 August 22

doi:10.1093/mnras/stx2156

A probabilistic approach to emission-line galaxy classification

R. S. de Souza,[1,2,3⋆] M. L. L. Dantas,[3⋆] M. V. Costa-Duarte,[3,4] E. D. Feigelson,[5] M. Killedar,[6] P.-Y. Lablanche,[7,8] R. Vilalta,[9] A. Krone-Martins,[10] R. Beck[11] and F. Gieseke[12]

[1]*Department of Physics and Astronomy, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3255, USA*
[2]*MTA Eötvös University, EIRSA "Lendulet" Astrophysics Research Group, Budapest 1117, Hungary*
[3]*Instituto de Astronomia, Geofísica e Ciências Atmosféricas, Universidade de São Paulo, R. do Matão 1226, 05508-090 São Paulo, Brazil*
[4]*Leiden Observatory, Leiden University, Niels Bohrweg 2, NL-2333 CA Leiden, the Netherlands*
[5]*Department of Astronomy and Astrostatistics and Center for Astrostatistics, Penn State University, 525 Davey Laboratory, University Park, PA 16802, USA*
[6]*Burnet Institute, 85 Commercial Road, Melbourne, VIC 3004, Australia*
[7]*Stellenbosch University, Matieland 7602, South Africa*
[8]*African Institute for Mathematical Sciences, 6 Melrose Road, Muizenberg, Cape Town 7945, South Africa*
[9]*Department of Computer Science, University of Houston, 4800 Calhoun Rd, Houston, TX 77204-3010, USA*
[10]*CENTRA/SIM, Faculdade de Ciências, Universidade de Lisboa, Ed. C8, Campo Grande, P-1749-016 Lisboa, Portugal*
[11]*Department of Physics of Complex Systems, Eötvös Loránd University, Budapest 1117, Hungary*
[12]*Radboud University Nijmegen, Toernooiveld 212, NL-6525 EC Nijmegen, the Netherlands*

Accepted 2017 August 18. Received 2017 August 17; in original form 2017 March 27

ABSTRACT

We invoke a Gaussian mixture model (GMM) to jointly analyse two traditional emission-line classification schemes of galaxy ionization sources: the Baldwin–Phillips–Terlevich (BPT) and $W_{\mathrm{H}\alpha}$ versus [N II]/Hα (WHAN) diagrams, using spectroscopic data from the Sloan Digital Sky Survey Data Release 7 and SEAGal/STARLIGHT data sets. We apply a GMM to empirically define classes of galaxies in a three-dimensional space spanned by the log [O III]/Hβ, log [N II]/Hα and log EW(Hα) optical parameters. The best-fitting GMM based on several statistical criteria suggests a solution around four Gaussian components (GCs), which are capable to explain up to 97 per cent of the data variance. Using elements of information theory, we compare each GC to their respective astronomical counterpart. GC1 and GC4 are associated with star-forming galaxies, suggesting the need to define a new starburst subgroup. GC2 is associated with BPT's active galactic nuclei (AGN) class and WHAN's weak AGN class. GC3 is associated with BPT's composite class and WHAN's strong AGN class. Conversely, there is no statistical evidence – based on four GCs – for the existence of a Seyfert/low-ionization nuclear emission-line region (LINER) dichotomy in our sample. Notwithstanding, the inclusion of an additional GC5 unravels it. The GC5 appears associated with the LINER and passive galaxies on the BPT and WHAN diagrams, respectively. This indicates that if the Seyfert/LINER dichotomy is there, it does not account significantly to the global data variance and may be overlooked by standard metrics of goodness of fit. Subtleties aside, we demonstrate the potential of our methodology to recover/unravel different objects inside the wilderness of astronomical data sets, without lacking the ability to convey physically interpretable results. The probabilistic classifications from the GMM analysis are publicly available within the COINtoolbox at https://cointoolbox.github.io/GMM_Catalogue/.

Key words: methods: data analysis – galaxies: evolution – galaxies: general – galaxies: nuclei – galaxies: star formation.

1 INTRODUCTION

Classification of objects has long been recognized as a major driver in natural sciences, from taxonomical classification of species, anthropological variation of cultures (e.g. Stocking 1968), to the

⋆E-mail: drsouza@ad.unc.edu (RSdeS); maria.luiza.dantas@usp.br (MLLD)

vastness of galaxy shapes (De Vaucouleurs 1959). Empirical classifications are powerful triggers for novel theories, an archetypal example being the Linnaean classification of organisms (Linnaeus 1758) that subsequently inspired the birth of Darwin's renowned theory of common descent (Darwin 1859).

Even though the properties of objects in nature may lie along a continuum, and groups may be defined by fuzzy boundaries, it may still be practical to divide them into categories that ideally reflect some physical distinctions. In astronomy, a canonical example is the one-dimensional Morgan–Keenan (Morgan & Keenan 1973) system of spectral stellar classification, in which stars of each class share similar ionization states or effective temperatures. The system was later used to compose the two-dimensional Hertzsprung–Russell diagram (e.g. Chiosi, Bertelli & Bressan 1992, and references therein), in which different stages of stellar evolution (e.g. main sequence, white dwarfs, giants, etc.) are grouped according to their luminosity (or magnitude) and effective temperature (or colour).

In the context of extragalactic astrophysics, various classification schemes have been proposed to help ascertain the main drivers regulating galaxy evolution; this task becomes imperative in the face of the deluge of information gathered by current (e.g. Sloan Digital Sky Survey, SDSS; York et al. 2000; Zhang & Zhao 2015) and upcoming (e.g. Large Synoptic Survey Telescope; Ivezic et al. 2009) large-scale sky surveys. Some examples are the classification of galaxies based on their morphological type (Lintott et al. 2008), their surrounding environment (von der Linden et al. 2010) or their spectral features (Morgan & Mayall 1957; Ucci et al. 2017).

Notably, the collisionally excited emission lines are powerful diagnostics to differentiate galaxies according to their ionization power source (e.g. Stasińska 2007), i.e. nuclear emission, star formation and so forth. Some of the most widely used emission-line diagnostics are the Baldwin–Phillips–Terlevich (BPT; Baldwin, Phillips & Terlevich 1981; Veilleux & Osterbrock 1987; Rola, Terlevich & Terlevich 1997; Kewley et al. 2001; Kauffmann et al. 2003; Stasinska et al. 2006; Schawinski et al. 2007) and, more recently, the $W_{\mathrm{H}\alpha}$ versus [N II]/Hα (WHAN; Cid Fernandes et al. 2010, 2011) diagrams.

The BPT diagram[1] classifies galaxies into star-forming (SF), composite and active galactic nuclei (AGN) hosts. The latter can be further subdivided into low-ionization nuclear emission-line region (LINER) galaxies and Seyferts. The lines used to define such classification are Hβ, [O III] λ5007, Hα and [N II] λ6583, and the galaxies are classified in the parameter space formed by log [N II]/Hα (x-axis) and log [O III]/Hβ (y-axis). SF galaxies are those in which the photoionization processes responsible for the emission lines are mainly due to young hot stars; they reside mostly in the left wing locus of the BPT diagram. On the opposite side lie the AGN-dominated objects, which are composed of two large groups, the LINERs and the Seyferts, usually divided by the line proposed by Schawinski et al. (2007). Objects classified as LINERs have an uncertain source of photoionization (Belfiore et al. 2016), which could be due to true nuclear activity or perhaps evolved stellar populations (Singh et al. 2013). Finally, we have the composite area of the diagram, marking the transition between SF and AGN objects, which are usually delimited by the theoretical extreme starburst line described by Kewley et al. (2001) and the empirical starburst line proposed by Kauffmann et al. (2003). It is worth noting that these boundaries are still a matter of debate and further alternative lines have been proposed (for instance Stasinska et al. 2006).

[1] Also known as the *Seagull* diagram.

The WHAN diagram has the same emission line ratio (i.e. log [N II]/Hα) on the x-axis as the BPT. On the other hand, it uses the equivalent width of Hα, i.e. log EW(Hα), as the characteristic parameter on the y-axis, instead of log [O III]/Hβ. The WHAN diagram uses a set of perpendicular and parallel straight lines to divide galaxies into strong AGN (sAGN), weak AGN (wAGN), SF and retired/passive galaxies. Because Hα is also used for the y-axis, a larger number of galaxies may be analysed, many of which would not appear on the BPT due to the lack of some emission features (Hβ and [O III] λ5007), i.e. groups of galaxies mainly represented by the *retired* and *passive* galaxy classes (Cid Fernandes et al. 2010, 2011; Stasińska et al. 2015). However, it lacks the definition of a transitional *composite* region.

Other examples of emission-line diagrams are the mass–excitation diagram (Juneau et al. 2014), which also uses the stellar mass of galaxies as a proxy for classification; the colour–excitation diagram (Yan et al. 2011); the blue diagram (Lamareille et al. 2004; Lamareille 2010) and the Trouille–Barger–Tremonti diagram (Trouille, Barger & Tremonti 2011). Moreover, many of these classification methods also include photometric information, such as the mid-infrared colour–colour diagrams (Lacy et al. 2004; Sajina, Lacy & Scott 2005; Stern et al. 2005). More recently, classifications based on ultraviolet (UV) information have been proposed in order to better understand the nature of galaxies at high redshifts (Feltre et al. 2016).

A common characteristic of most of these diagrams and the majority of standard classification systems in astronomy is the sharp division between classes, in which boundaries are more often than not defined by eye or fitted without accounting for a smooth transition between objects. Given the ever-increasing richness of information enclosed in astronomical surveys, we advocate updating standard classification schemes under the paradigm of contemporary statistical methods, while still maintaining the crucial role of expert knowledge in the physical interpretation of data-driven classes.

From a methodological point of view, a recent trend in object classification has been the reliance on *machine learning* for data analysis (Hastie, Tibshirani & Friedman 2009; Murphy 2012). While being conceptually very similar to well-known existing statistical methods, the tremendous increase in both available data and computational power over the past two decades has led to the development of a variety of advanced techniques. Machine learning aims at deriving 'models' that can retrieve useful information in an automatic manner. Examples of machine learning in astronomy are supernovae classification (e.g. Richards et al. 2012; Ishida & de Souza 2013; Karpenka, Feroz & Hobson 2013; Lochner et al. 2016; Sasdelli et al. 2016), studies of emission-line spectra of galaxies (Beck et al. 2016; Ucci et al. 2017), photometric redshift estimation (e.g. Collister & Lahav 2004; Krone-Martins, Ishida & de Souza 2014; Cavuoti et al. 2015; Elliott et al. 2015; Hogan, Fairbairn & Seeburn 2015; Beck et al. 2017) and detection of galaxy outliers (e.g. Baron & Poznanski 2017).

In the past few decades, various new techniques have been proposed along two prominent lines of research: supervised and unsupervised learning (Hastie et al. 2009). For the former, one is given labelled data, i.e. objects described by a set of parameters along with a class label (e.g. discrimination between early- and late-type galaxies based on their photometric colours). The other line of research, unsupervised learning, does not make assumptions about pre-existing labels; the goal is to automatically derive conclusions about the data structure by assigning 'similar' objects to the same group. Thus, in contrast to supervised classification, no information is made available about the categories or classes to which

objects belong. Instead, the unsupervised learning model must discover such classes. While supervised learning methods have been applied previously to the specific problem of AGN classification (Beck et al. 2016), the unsupervised approach is by definition more suitable for challenging or reinforcing the existing classification paradigm, as it can study what statistical evidence the measurement data contain in support of given classes. For this reason, in this paper we adopt an unsupervised approach.

One type of unsupervised model is the so-called Gaussian mixture model (GMM; e.g. Everitt et al. 2011). Most unsupervised clustering methods, like the popular friends-of-friends algorithm (also known as single-linkage agglomerative clustering in statistical parlance; Davis et al. 1985), are non-parametric and less robust against different choices of algorithms (Feigelson & Babu 2012). GMMs, in contrary are parametric, hence solvable by maximum likelihood approach. This makes the GMMs a *desideratum* due to its objective, stable and interpretable probabilistic results.

Previous examples of the application of mixture models in astronomy are the search for subcluster structures of young stars in massive star-forming regions (Kuhn et al. 2014), and the separation of millisecond pulsars from a broader sample (Lee et al. 2012). This paper demonstrates the application of GMM for the emission-line classification of galaxies, and further discusses how the data-driven groups can be related to classic classifications, which are based on expert domain knowledge.

The outline of this paper is as follows. In Section 2 we provide an overview of the sample selection. Section 3 describes the GMM methodology. We present our main results in Section 4, discuss their physical meaning in Sections 5 and 6 and present our conclusions in Section 8. The standard Λ cold dark matter (ΛCDM) cosmology with $\{H_0, \Omega_M, \Omega_\Lambda\} = \{70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}, 0.3, 0.7\}$ has been used throughout the paper.

2 CATALOGUE

The galaxy sample used in this work is the result of matching two data bases: the SDSS Data Release 7 (DR7; Abazajian et al. 2009) and the public SEAGal/STARLIGHT catalogue.[2] The SDSS-DR7 comprises photometry in five broad-band filters (*ugriz*) and optical spectroscopy between 3800 and 9200 Å in the observed frame. Our initial sample retrieved from the SDSS-DR7 data base is volume limited[3] and composed of galaxies brighter than $M_r < -19.88 + 5\log h_{70}$, with $h_{70} \equiv H_0/70\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$, over the redshift range $0.015 < z < 0.075$.[4] The magnitudes in our sample were treated to account for the effects of Galactic extinction, and further *K*-corrected using the software KCORRECT v3.2 (Blanton et al. 2003a).

The SEAGal/STARLIGHT catalogue provides spectral synthesis parameters such as average stellar metallicity ($\langle Z/Z_\odot \rangle_L$, with respect to the Sun's metallicity), average stellar age ($\langle \log(t/\mathrm{yr}) \rangle_L$, in units of year) and the 4000 Å break (D_n4000),[5] as well as emission-line measurements of all SDSS-DR7 galaxies. The empirical spectral synthesis technique is carried out using the STARLIGHT code (Cid Fernandes et al. 2005), which fits the stellar continuum by using a library of simple stellar populations from Bruzual & Charlot (2003).

The emission lines are fitted by subtracting the stellar continuum and using a Gaussian profile (for more details, see Cid Fernandes et al. 2010). For this analysis, fluxes and equivalent widths of the [O III] λ5007, H β, H α and [N II] λ6583 emission lines are extracted from the SEAGal/STARLIGHT data base. In order to ensure good quality measurements, we impose a signal-to-noise ratio (S/N) of >3 and a fraction of bad pixels lower than 25 per cent, for all emission lines. The aforementioned constraints and matching between both data bases lead to a final galaxy sample that consists of 83 578 objects. Fig. 1 displays the projections of all galaxies in the sample on the traditional BPT and WHAN diagrams.

3 GAUSSIAN MIXTURE MODELS

GMM is a parametric model that, within a given feature space, assumes the existence of classes that can be described by a superposition of multivariate Gaussian distributions (e.g. McLachlan & Peel 2000; Hastie, Tibshirani & Friedman 2001; Mengersen, Robert & Titterington 2011; Murphy 2012). The model is defined as a probability density function composed of a weighted summation of Gaussian component (GC) densities. The goal is to describe the distribution of data in a certain feature space and assign probabilities to the membership of a given datum in each class. More specifically, for a total of K clusters in a d-dimensional parameter space, the GMM is a probability distribution $p(x)$ given by a weighted summation of K components :

$$p(x) = \sum_{\kappa=1}^{k} \zeta_k \phi(x; \boldsymbol{\mu}_\kappa, \boldsymbol{\Sigma}_\kappa), \tag{1}$$

with mixture weights denoted by ζ_κ, and $\sum \zeta_\kappa = 1$. Here, each of the K model components is described as a d-variate Gaussian density, fully characterized by its mean $\boldsymbol{\mu}_\kappa$ and covariance matrix $\boldsymbol{\Sigma}_\kappa$:

$$\phi(x; \boldsymbol{\mu}_\kappa, \boldsymbol{\Sigma}_\kappa) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_\kappa|}} \, \mathrm{e}^{-\frac{1}{2}(x-\boldsymbol{\mu}_\kappa)\boldsymbol{\Sigma}_\kappa^{-1}(x-\boldsymbol{\mu}_\kappa)}. \tag{2}$$

Various ways exist to fit such a GMM to a given set of data points, among which stands out the popular expectation–maximization (EM) algorithm (Dempster, Laird & Rubin 1977; McLachlan & Krishnan 2008). This work adopts the EM algorithm from the R (R Core Team 2016) package MCLUST (Fraley & Raftery 2002) to fit the GMMs.[6]

4 GMM APPLICATION TO SDSS AND SEAGAL DATA

This section presents the results from the application of a GMM to the SDSS-DR7 and SEAGal/STARLIGHT catalogues.

4.1 Results

We now apply the GMM to our galaxy catalogue projected into the joint combination of the BPT and WHAN diagrams. Hence, the

[2] http://casjobs.starlight.ufsc.br/casjobs/

[3] This choice follows a similar procedure as e.g. Mateus et al. (2006) and Cid Fernandes et al. (2011), and aims to mitigate the bias towards galaxies with the presence of strong emission lines.

[4] The narrow redshift range herein employed has the aim to mitigate potential biases caused by evolutionary effects.

[5] For more information, we refer the reader to the STARLIGHT CasJobs Schema Browser: http://casjobs.starlight.ufsc.br/casjobs/field_list.html

[6] Additionally, an independent GMM was implemented using the PYTHON package SCIKIT-LEARN (Pedregosa et al. 2011) to check the cross-consistency of our results.

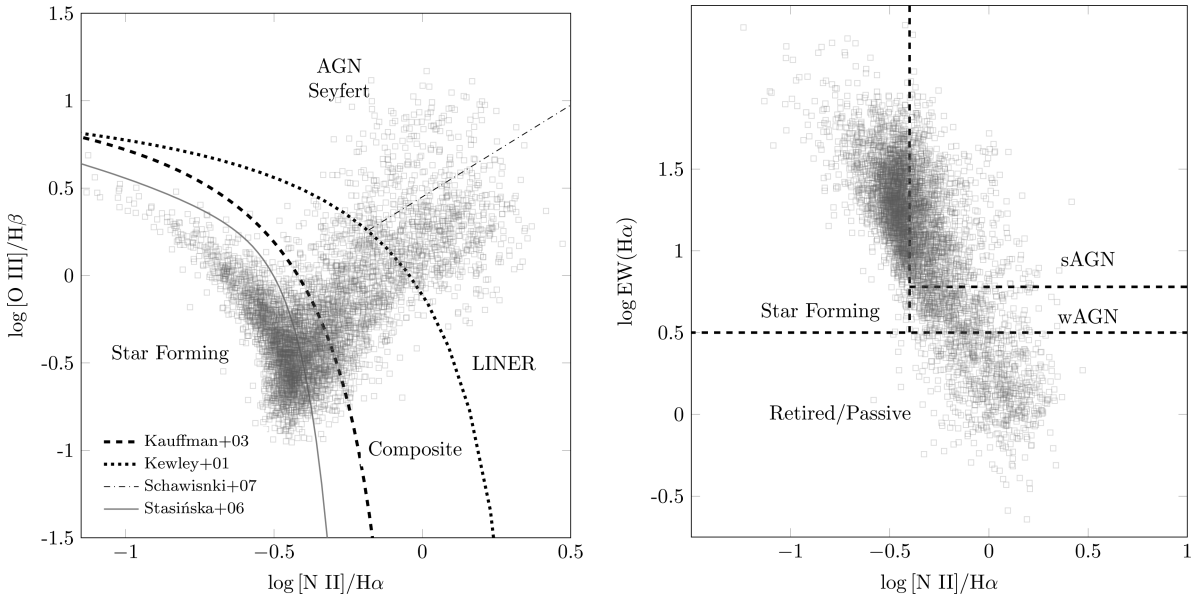

Figure 1. BPT and WHAN diagrams, from left to right, with galaxy points from the SDSS and SEAGal data sets. On the BPT diagram, the curves define the division between SF and AGN classes (dotted: Kewley et al. 2001; solid: Stasinska et al. 2006; dashed: Kauffmann et al. 2003), and the dot–dashed line shows the division between AGN and LINERs as suggested by Schawinski et al. (2007). On the WHAN diagram, the dashed straight lines discriminate between sAGN, wAGN, SF and retired/passive galaxies (Cid Fernandes et al. 2011). For better visualization, the points in the figure represent a subsample of 10 000 randomly selected galaxies.

dimension of the parameter space is $d = 3$ and the data vector $\boldsymbol{x}$ in equation (1) is given by

$$\boldsymbol{x} = \begin{pmatrix} \log\,[\mathrm{N\,II}]/\mathrm{H}\alpha \\ \log\,[\mathrm{O\,III}]/\mathrm{H}\beta \\ \log\,\mathrm{EW}(\mathrm{H}\alpha) \end{pmatrix}. \quad (3)$$

The output is a soft classification of each object given by the membership probability for each group, together with parameters $\boldsymbol{\mu}_\kappa$ and $\boldsymbol{\Sigma}_\kappa$ for each three-variate GC.

We show the results in Fig. 2 of the two-, three- and four-cluster solutions, each projected on to the two-dimensional BPT and WHAN parameter space. A visual inspection suggests that while the whole population of galaxies cannot be explained by a single Gaussian distribution, their overall distribution can be approximated by multiple Gaussian clusters. The contours represent 68 and 95 per cent confidence levels around the mean of each GC, respectively. The left-hand panel of Fig. 2 shows that the two-cluster solution roughly separates the SF and AGN-dominated galaxies in both diagrams. The solution with three clusters, displayed in the middle panel, identifies a composite region of the BPT diagram, and a possible transitional region in the WHAN diagram, which will be further discussed in Section 5. The solution with four clusters indicates a possible subdivision of the SF region in the BPT, which may be connected to the existence of starburst galaxies, predominantly located in the top-left region of the BPT diagram. The parameters for the four-cluster solution are shown in Table 1. Four GCs are preferred to describe the galaxy population in the log [N II]/H α, log [O III]/H β and log EW(H α) feature space, based on a set of cluster validation methods, as described next.

4.2 Internal cluster validation

Cluster validation plays a key role in assessing the quality of a given clustering structure. It is called internal when statistics are devised to capture the quality of the induced clusters using solely the available data objects. Four validation measures are used: Bayesian information criterion (BIC; Schwarz 1978), but see also Drton & Plummer (2017), integrated complete likelihood (ICL; Biernacki, Celeux & Govaert 2000), entropy (Baudry et al. 2010) and silhouette (Rousseeuw 1987) diagnostics. See Appendix A for details of each of the former methodologies. Below, the scrutiny of the adopted diagnostics for GMM solutions up to 10 GCs.

BIC and ICL solutions are shown on the left-hand panel of Fig. 3. Note that BIC fails to constrain the model to a reasonably low number of groups. ICL on the other hand suggests a lower number of GCs. Similar differences between BIC and ICL are well known in the statistical literature (e.g. Biernacki et al. 2000). Entropy values for solutions ranging from 2 to 10 clusters are shown in the middle panel of Fig. 3. There is an elbow in the plot at $K = 3$ GCs, which, together with the preference of ICL, leads us to focus our attention around this solution. Additionally, the silhouette values are displayed in the right-hand panel of Fig. 3, suggesting the preference for only two groups.

At this point, multiple standard internal validation methods suggest that 2–3 clusters are present in the data, but the discrimination is not clear and more cluster components are compatible with the internal validation measurements. In the following, we propose the use of a residual analysis to *break the tie* and quantify how well each model performs in terms of synthetically reproducing the original data structure.

4.3 Residual analysis

Residual analysis is one of the most informative methods to check a model fit. It helps to measure how well a statistical model explains the data at hand and its ability to predict future sets of observations (see e.g. Lindsay & Roeder 1992; Cui et al. 2015, for applications of residual analysis in the context of mixture models).

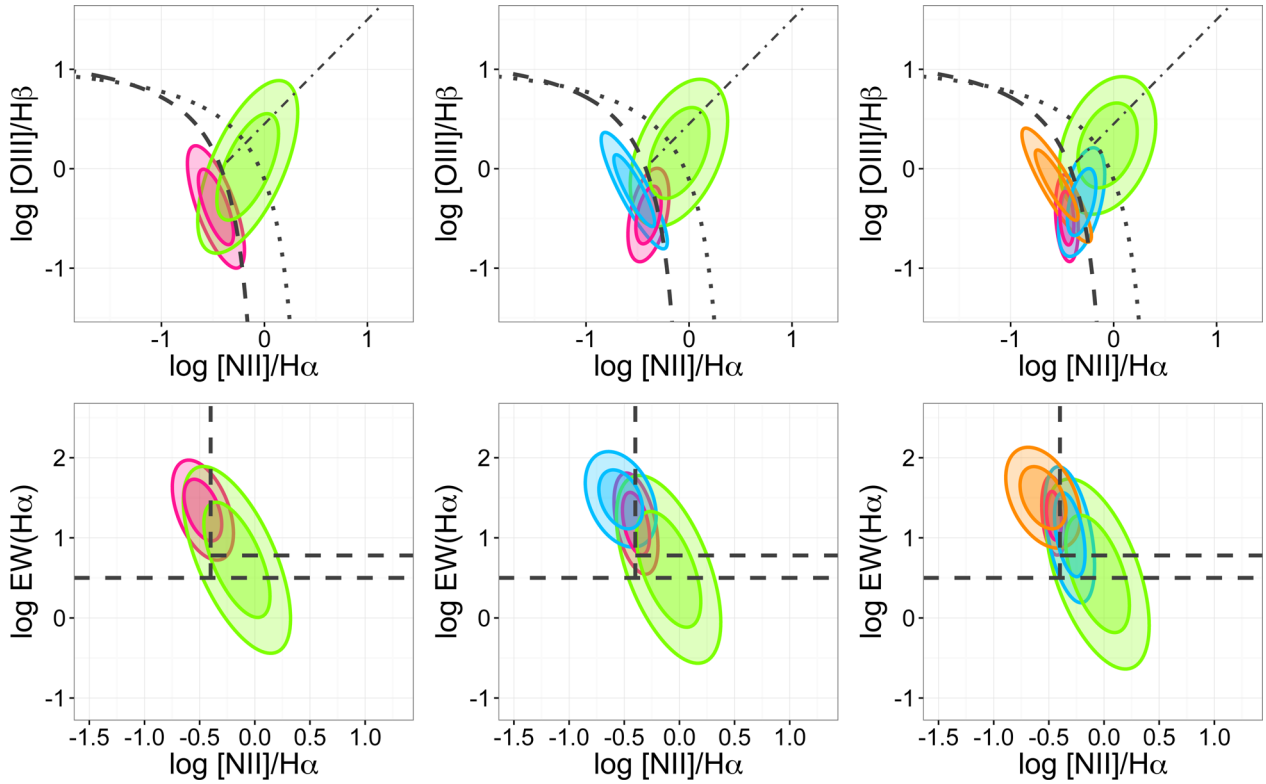

Figure 2. The GCs projected on to the BPT (top panels) and WHAN (bottom panels) diagrams. From left to right are the solutions for two, three and four GCs. For each component the thick lines represent 68 and 95 per cent confidence levels, respectively.

Table 1. Parameters of the GMM solution with four GCs for the galaxy distribution in the log [N II]/H α, log [O III]/H β and log EW(H α) space. Shown are the mixture weights $\zeta_1 \ldots \zeta_4$, central vectors of the clusters, $\boldsymbol{\mu}_1 \ldots \boldsymbol{\mu}_4$, and covariance matrices $\boldsymbol{\Sigma}_1 \ldots \boldsymbol{\Sigma}_4$.

Parameter	Value
ζ_1	0.281
ζ_2	0.252
ζ_3	0.276
ζ_4	0.189
$\boldsymbol{\mu}_1$	$(-0.454 \quad -0.497 \quad 1.276)$
$\boldsymbol{\mu}_2$	$(-0.058 \quad 0.234 \quad 0.549)$
$\boldsymbol{\mu}_3$	$(-0.310 \quad -0.335 \quad 1.039)$
$\boldsymbol{\mu}_4$	$(-0.552 \quad -0.165 \quad 1.501)$
$\boldsymbol{\Sigma}_1$	$\begin{pmatrix} 2.04\times10^{-3} & -1.93\times10^{-3} & -2.41\times10^{-3} \\ -1.93\times10^{-3} & 3.18\times10^{-2} & -5.16\times10^{-3} \\ -2.41\times10^{-3} & -5.16\times10^{-3} & 4.04\times10^{-2} \end{pmatrix}$
$\boldsymbol{\Sigma}_2$	$\begin{pmatrix} 3.65\times10^{-2} & 1.68\times10^{-2} & -4.99\times10^{-2} \\ 1.68\times10^{-2} & 7.99\times10^{-2} & 9.57\times10^{-3} \\ -4.99\times10^{-2} & 9.57\times10^{-3} & 2.35\times10^{-1} \end{pmatrix}$
$\boldsymbol{\Sigma}_3$	$\begin{pmatrix} 8.42\times10^{-3} & 1.04\times10^{-2} & -1.45\times10^{-2} \\ 1.04\times10^{-2} & 4.98\times10^{-2} & -5.07\times10^{-2} \\ -1.45\times10^{-2} & -5.07\times10^{-2} & 1.21\times10^{-1} \end{pmatrix}$
$\boldsymbol{\Sigma}_4$	$\begin{pmatrix} 1.90\times10^{-2} & -2.87\times10^{-2} & -1.42\times10^{-2} \\ -2.87\times10^{-2} & 5.49\times10^{-2} & -2.74\times10^{-2} \\ -1.42\times10^{-2} & 2.74\times10^{-2} & 6.62\times10^{-2} \end{pmatrix}$

In order to check the goodness of fit of each GMM solution, a comparison between the synthetic and observed data for each model projected on to the BPT and WHAN diagrams is performed. Results are presented in Fig. 4, which shows the smoothed observed data contrasted with the GMM solutions for two, three and four GCs.[7] A visual analysis of Fig. 4 reveals that on the BPT diagram, the solution with two clusters barely reproduces the two-wing shape, and the four-cluster solution seems to be a nearly perfect match with the original data. On the WHAN diagram a visual inspection does not lead to equally clear conclusions, but the solution with four groups seems to be preferred.

A quantitative analysis of Fig. 4 is displayed in Fig. 5, which shows the residual map for each solution together with a linear fit between the smoothed observed and simulated data for each GMM solution. Note that for the BPT diagram, each increase in the number of GCs considerably improves the amount of variance explained by the model, which is consistent with the visual analysis of Fig. 4. The solution with four GCs is able to explain up to 97 per cent of the data variance. For the WHAN diagram the distinction between the solutions with two and three clusters is fuzzy, but the solution with four GCs is equally capable of explaining 97 per cent of the data variance. Combining the residual analysis and the internal validation methods previously described leads us to keep the solution with four GCs as our 'fiducial model' hereafter.

[7] To smooth the residual maps, we use kernels with a bandwidth of 0.05 within a grid of 100 × 100.

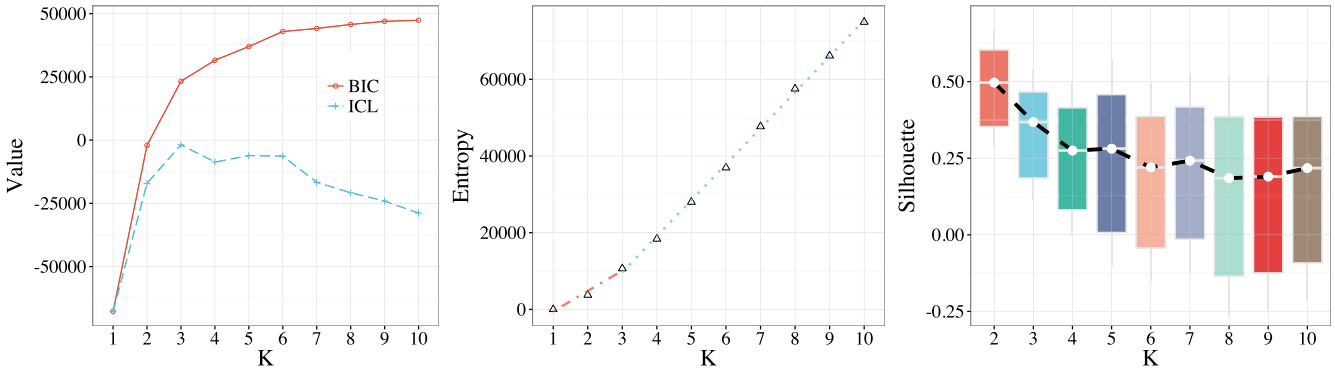

Figure 3. Results for K number of GCs for each of the internal validation methods as follows: the left-hand panel shows the BIC and ICL values, the middle panel shows the entropy elbow diagnostics and the right-hand panel shows the silhouette results.

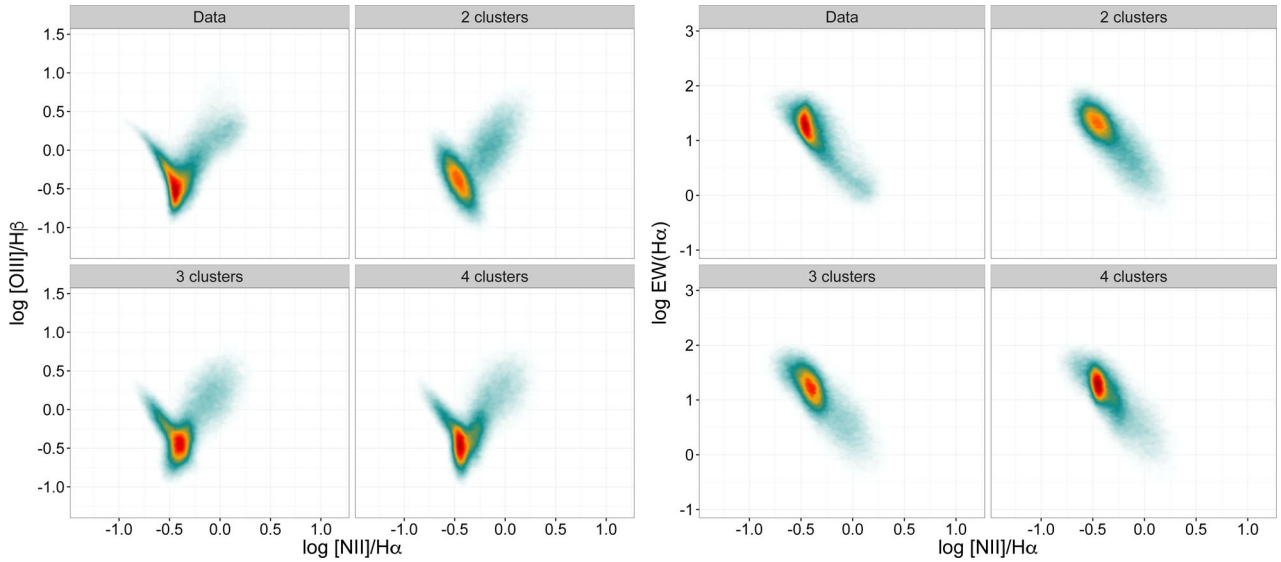

Figure 4. Comparison between observed and synthetic data for two-, three- and four-cluster solutions on the BPT and WHAN diagrams. Both the original and synthetic data are smoothed using the same kernel.

5 EXTERNAL CLUSTER VALIDATION APPLIED TO THE GMM SOLUTION

This section discusses how to attribute physical meaning to the statistically motivated groups, and provide the means on how to compare them to current classification schemes. If the cluster validation is performed against an external and independent classification of objects (e.g. the BPT and WHAN classifications), the validation is called external. External cluster validation (ECV) is based on the assumption that an understanding of the output of the clustering algorithm can be achieved by finding a resemblance of the clusters to existing classes. In the present application, ECV is used to compare the cluster structure produced by a GMM to the class structure corresponding to well-established galaxy classification schemes.

The idea is to use an objective methodology to decide if the induced clusters have recovered an existing classification, or if there is evidence to claim the existence of novel data groups not resembling existing classes. Specifically, we use a probabilistic approach to individually compute the distance between each cluster and its most similar class; the degree of separation between cluster and class can then be analysed to decide on the scientific value behind a small distance (near match) or long distance (strong disagreement).

The methodology herein employed follows the work of Vilalta, Stepinski & Achari (2007); we briefly describe its main concepts in Appendix B. The method relies on the estimate of the Kullback–Leibler (KL) distance (a measure of relative entropy; Kullback & Leibler 1951) between different groups projected into one dimension via linear discriminate analysis (LDA). Smaller KL distances are found for closer groups.

To illustrate the application of ECV methodology in our data set, we show a pairwise comparison of the one-dimensional linear discriminant projections of the four GCs to the BPT and WHAN classifications in Figs 6 and 7, respectively. The figure depicts each group – colour coded as in Fig. 2 – alongside the astrophysically motivated classes.

A visual inspection of Figs 6 and 7 reveals that groups GC1 and GC4 are closer to SF galaxies in both BPT/WHAN diagrams, while GC2 and GC3 are closer to AGN/(wAGN and retired/passive) and composite/sAGN, respectively. Thus, there is no particular evidence for a new group, but surprisingly GMMs are capable of automatically identifying groups of galaxies resembling the traditional classification scheme of both diagrams from a higher dimensional feature space. Table 2 summarizes the results showing each class and its closest GC alongside to $1 - \mathrm{KL}$ distance.

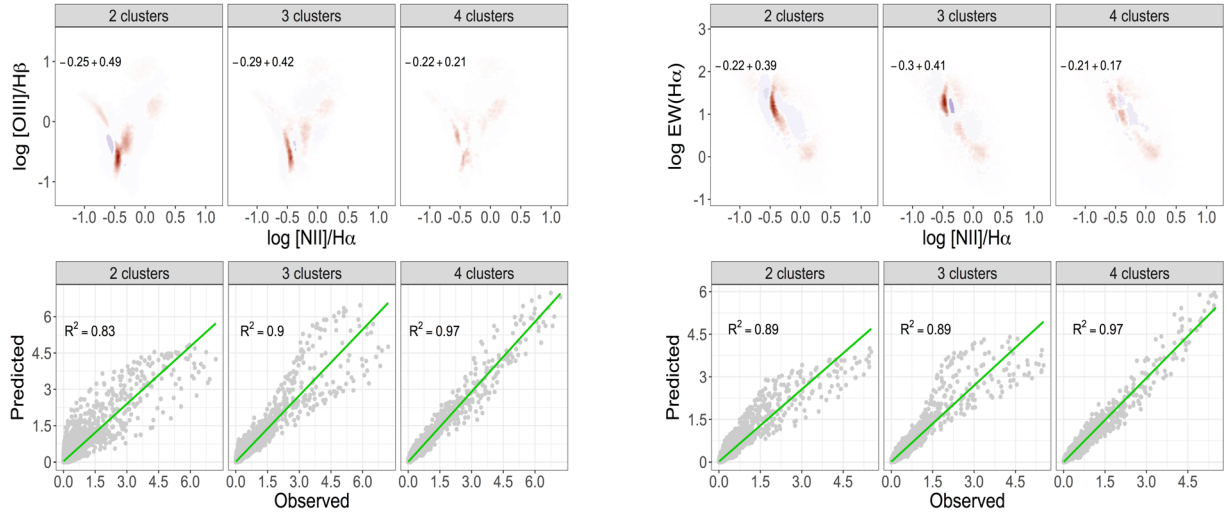

Figure 5. Goodness of fit diagnostics for the two, three and four GCs solutions projected on to the BPT (left-hand panel) and WHAN (right-hand panel) diagrams. Top: residual surface density, negative residuals are shown in blue and positive residuals are shown in red. Also displayed is the maximum variation of the residual map in comparison to the original data. Note that for the solution with four GCs, the maximum difference between the simulated and original data is always $\lesssim$22 per cent. Bottom: surface density of the mixture model solution is plotted against surface density of the smoothed observed data. A linear fit of predicted versus observed values is green, and on the left-hand side of each panel we indicate the proportion of variance explained, R^2.

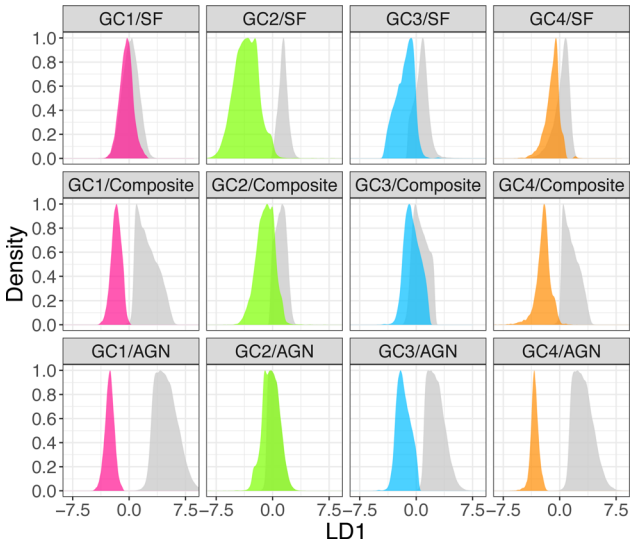

Figure 6. Density distributions in the one-dimensional linear discriminant projections for each of the four GCs (coloured distributions) compared to the traditional BPT classification of SF, composite and AGN galaxies (grey distributions).

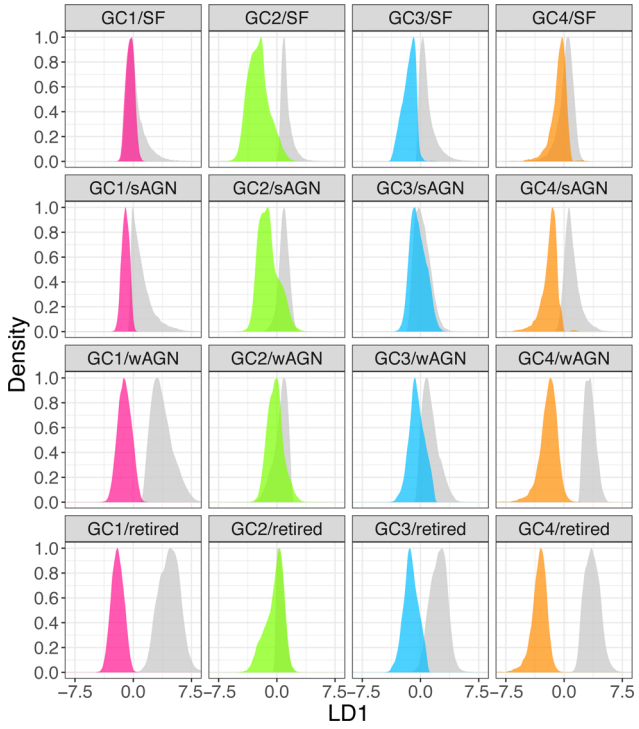

Figure 7. Density distributions in the one-dimensional linear discriminant projections for each of the four GCs (coloured distributions) compared to the traditional WHAN classification of SF, sAGN, wAGN and retired galaxies (grey distributions).

In order to better visualize the closest associations (i.e. with a normalized KL $\lesssim$ 0.05), we show a chord diagram (Gu et al. 2014; De Souza & Ciardi 2015) in Fig. 8. It illustrates the level of relationship between distinct groups, which are represented by segments around the circle. Normalized distances between distributions are shown as ribbons; the thickness of the ribbons is weighted by 1 − KL distance between each pair of groups, so the thicker the ribbon, the closer the GC to its traditional classification counterpart.

5.1 GC1/GC4 interpretation

The connection between GC1, GC4 and the SF region in both BPT and WHAN diagrams is straightforward and somewhat expected. In terms of the WHAN diagram, by construction (Cid Fernandes et al. 2011), the vertical line at log [N II]/Hα-0.40 represents an optimal transposition of the Stasinska et al. (2006) SF/AGN-BPT division projected into the WHAN diagram. Consequently, the combination of both diagrams in a three-dimensional space should still preserve the same locus for SF-dominated galaxies, which is automatically retrieved by the GMM methodology.

Albeit the solution with three GCs (90 per cent of the data variance) only requires a single GC within the SF region, the solution

Table 2. Summary of the associations between GMM, BPT and WHAN groups. Next to each class is one minus the KL distance of each BPT and WHAN class to its respective GC.

GMM	BPT	WHAN
GC1	Star forming (0.981)	Star forming (0.996)
GC2	AGN (0.981)	wAGN(0.961)+retired (0.983)
GC3	Composite (0.943)	sAGN (0.984)
GC4	Star forming (0.934)	Star forming (0.957)

with four GCs (97 per cent of the data variance) splits the SF region into GC1 and GC4, a behaviour that may be physically interpreted by the presence of starburst galaxies, predominantly populating the top-left wing of the BPT diagram. To be more specific, galaxies at the top-left wing have current specific SFRs about two orders of magnitude larger than the metal-rich galaxies at its bottom (see fig. 2 in Asari et al. 2007).

5.2 GC2 interpretation

The locus occupied by the GC2 in the three-dimensional emission-line space is concomitantly connected to the BPT-AGN region and mainly wAGN+retired region (seconded by the sAGN region) in the WHAN diagram. The result may appear controversial at first glimpse, as one could expect that the BPT-AGN galaxies should relate to the sAGN galaxies in the WHAN diagram. Nonetheless, the GMM recovers a previous finding by Cid Fernandes et al. (2010). The authors show that the dichotomy between Seyferts and LINERs is wiped out by the presence of weak line galaxies in the sample, which are usually left out from vanilla emission-line galaxy studies solely relying on the BPT plane. They suggest that the right wing of the BPT diagram is actually populated by AGN and retired galaxies, which is corroborated by the GMM results. In other words, by finding a larger group composed by AGN-BPT and wAGN+retired-WHAN galaxies, our method does not show a statistical evidence for the separation between Seyferts and LINERs as independent subclasses.

5.3 GC3 interpretation

As aforementioned, Figs 6 and 7 indicate that GC2 also relates to the sAGN region at the WHAN diagram in a lesser extent than the GC3, which we shall discuss next. GC3 relates mostly to the composite-BPT and sAGN-WHAN regions. While it is desirable that GMM finds the composite-BPT locus, the connection to the sAGN-WHAN galaxies is not so straightforward. This can be explained due to the lack of a formal composite area in such diagram. From Fig. 2, we see that GC3 occupies a transitional region between GC1/GC4 and GC2; a locus that could also be designated as an 'effective composite' area between SF and sAGN-dominated galaxies.

6 AGE, METALLICITY AND 4000 Å BREAK DISTRIBUTIONS FOR THE GMM GROUPS

We now address whether these data-driven groups bring new insights beyond what is given by established classification schemes. This section discusses characteristics of the galaxies in each GC alongside the BPT and WHAN classes.

It is well known that different galaxy properties share some sort of *symbiotic* relationship; for instance, D_n4000 is closely linked to the characteristics of the stellar population (e.g. average population age, metallicity; as described in Poggianti & Barbaro 1997; Blanton et al. 2003b; Goto 2003; Costa-Duarte, Sodré & Durret 2013; Stasińska et al. 2015; Vazdekis et al. 2016, and references therein). Hence, the aforementioned features serve as proxies to derive other galaxy properties (e.g. star formation rate; Tinsley 1980; Zaritsky 1993; Poggianti & Barbaro 1997; Vazdekis et al. 2016).

In order to probe how the GCs compare to the classes derived by classical diagrams, we look into their properties not explicitly used in the GMM analysis. For that purpose, we choose three of the main features retrieved from the SEAGal/STARLIGHT output data, as defined in Section 2: $\langle Z/Z_{\odot}\rangle_L$, $\langle \log(t/\mathrm{yr})\rangle_L$ and D_n4000.[8] The goal is to check how these properties vary according to the employed classification: GMM, BPT and WHAN. To that end, we portray their statistical properties as boxplots in Figs 9–11, as well as their summary statistics in Table 3, which shows the values for the median, first (Q1) and third (Q3) quartiles, and the interquartile range (IQR ≡ Q3–Q1).

On the boxplots, the groups are vertically aligned based on their proximity in terms of the KL distance, and the fiducial order roughly follows increasing values of $\log$ [N II]/Hα in the BPT diagram (i.e. from left to right: SF, composite, AGN). As we can see from the boxplots, by aligning these groups in this way, a positive monotonic relationship between the classes and the median values of $\langle Z/Z_{\odot}\rangle_L$, $\langle \log(t/\mathrm{yr})\rangle_L$ and D_n4000 distributions is revealed. The trend is a consequence of the AGN host galaxies having different characteristics from their inactive counterparts, preferentially populating the so-called green valley and red sequence of the colour–mass diagram (e.g. Schawinski et al. 2010). Thus, as we move towards the right-hand side of the BPT diagram, one is mostly looking at early-type galaxies, that are characterized by older and more metallic stellar populations, and higher values of D_n4000 (e.g. Poggianti & Barbaro 1997; Schawinski et al. 2010; De Souza et al. 2016).

Notably, the GMM solutions automatically find groups that share meaningful physical properties, beyond the features used in the clustering algorithm. An inspection of Table 3 confirms that the statistical properties are quite similar between the GMM groups and their respective counterparts on the BPT and WHAN diagrams. Additionally, Table 3 shows that the distributions of galaxy properties in SF groups are consistent between the BPT and WHAN diagrams in terms of medians and IQR. Their values roughly lie between those of GC4 and GC1. For instance, GC4 and GC1 have median values for $\langle Z/Z_{\odot}\rangle_L$ of 0.50 and 0.58, while the BPT and WHAN SF groups have values of 0.56 and 0.55, respectively.

In the case of $\langle Z/Z_{\odot}\rangle_L$ and D_n4000, the median and IQR increase more steadily for the GCs, in comparison to the BPT and WHAN classes. The trend of D_n4000 visible in Fig. 11 indicates that different types of galaxies occupy different loci in the GMM classification. For instance, GC4, the first one on the left, has low median values for those parameters, which is in agreement with the characteristics of young stellar populations, i.e. SF galaxies. On the other hand, GC2, composed mostly of AGN hosts and retired/passive objects, has a higher median value of D_n4000, which is in accordance with older stellar populations. As D_n4000 can be used as proxy for morphology (Dressler & Gunn 1990; Brinchmann et al. 2004), the higher D_n4000 median highlights the AGN preference to reside in early-type galaxies (Schawinski et al. 2010; De Souza et al. 2016). Besides, the larger IQR for GC2, corroborates

[8] Note that we are using average values weighted by flux/luminosity due to the smaller uncertainties on those (see table 1 in Cid Fernandes et al. 2005). In the case of missing values, synthetic D_n4000 was used. This is done for a better sampling statistics, but it does not affect the overall results.

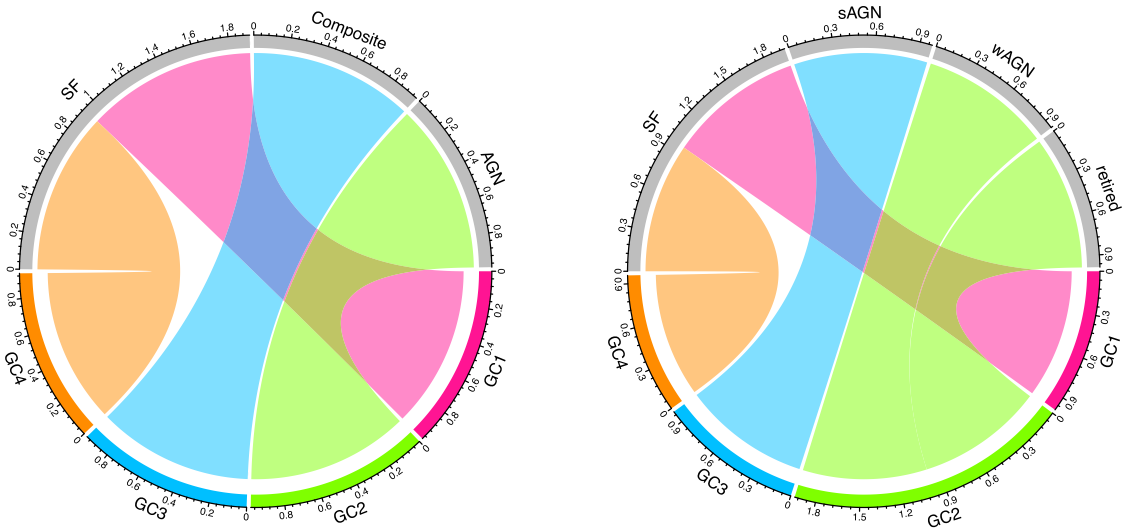

Figure 8. Chord diagrams representing the associations between the GCs and the astronomical classification classes defined on the BPT, left-hand panel, and the WHAN, right-hand panel, diagrams. A thicker connecting ribbon indicates stronger association.

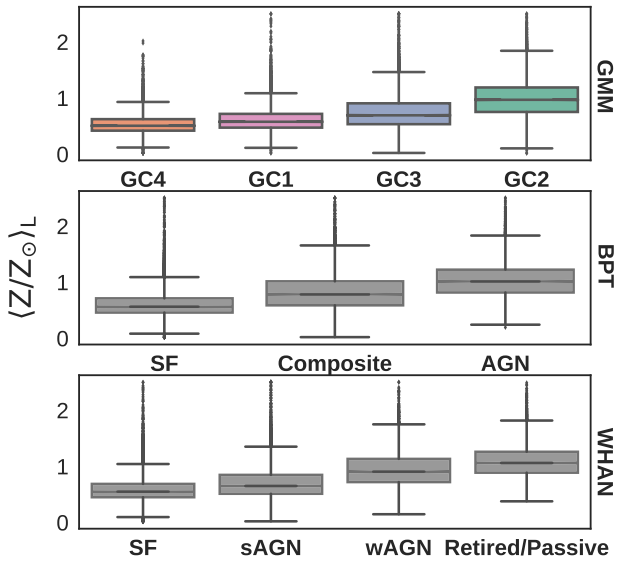

Figure 9. Metallicity distributions portrayed as boxplots for GMM, BPT and WHAN diagram classes, from top to bottom. For the GMM we can see the GCs displayed in the following order: GC4, GC1, GC3, GC2 following increasing log [N II]/H α, i.e. the x-axis of the BPT diagram. The order of the remaining groups is given by the KL distance to the GMM components. The width of boxes is proportional to the square root of the number of galaxies within each bin and the whiskers extend to the most extreme data point, which is within the 50 per cent interquartile range (IQR). To better illustrate the overall distribution of the sample, for the boxplots of the BPT diagram we have omitted part of the outlier zone, 'zooming' into the interquartile distance.

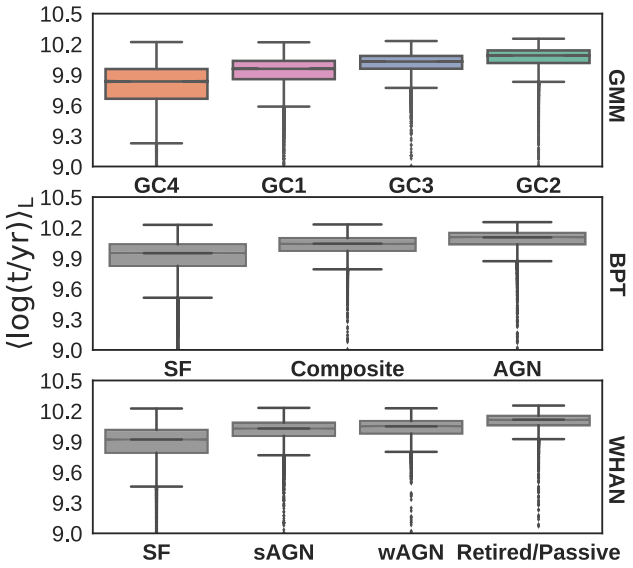

Figure 10. Average stellar age distributions portrayed as boxplots for GMM, BPT and WHAN diagram classes, from top to bottom. The GCs and remaining groups are ordered as in Fig. 9.

with the fact that AGN can reside either within early- or late-type galaxies. The decreasing trend in terms of median and IQR found for $\langle \log (t/\mathrm{yr}) \rangle_L$, within the GMM groups is overall consistent with the traditional diagrams as well.

The GMM systematically finds groups that have a sharper differentiation of their values of $\langle Z/Z_{\odot} \rangle_L$, $\langle \log (t/\mathrm{yr}) \rangle_L$ and D_n4000, when compared to the standard classification, especially in terms of distributing most of the dispersion into fewer clusters (usually only one), yielding to a majority of lower dispersion classes. These findings elucidate the power of the proposed method – since the physical parameters were not included in the GMM classification, their favourable behaviour within and between the GCs has been inherently caused by the method applied.

7 DIGGING DEEPER – THE SEYFERT/LINERS DICHOTOMY AND THE QUEST FOR PASSIVE GALAXIES

Internal validation methods present a trade-off between predictive power and simplicity. In other words, a good model should describe the data as best as possible with the fewer number of groups necessary. Whilst our fiducial model based on diverse criteria points for a solution around 3–4 groups, there is a physical motivation to look further and see if we can spot the presence of LINERs and discriminate the passive/retired galaxies in our sample.

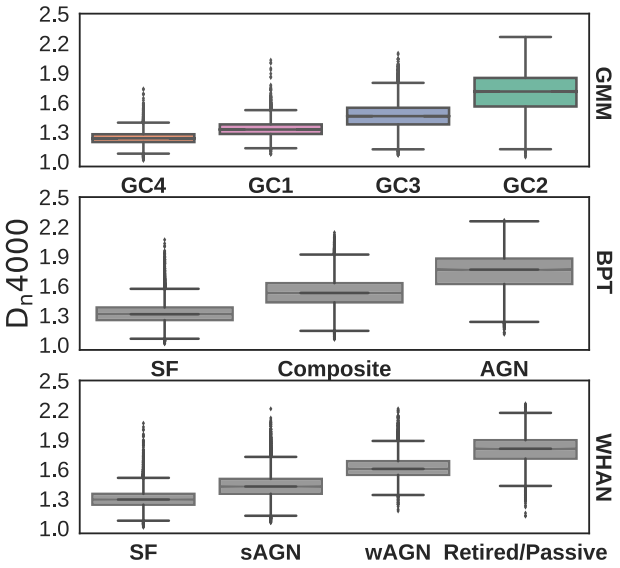

Figure 11. D_n4000 distributions portrayed as boxplots for GMM, BPT and WHAN diagram classes, from top to bottom. The GCs and remaining groups are ordered as in Fig. 9.

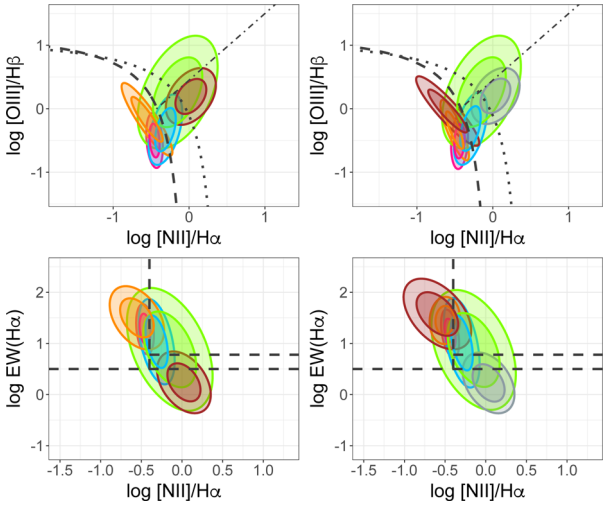

Figure 12. The GCs projected on to the BPT (top panels) and WHAN (bottom panels) diagrams. From left to right are the solutions for five and six GCs. For each component the thick lines represent 68 and 95 per cent confidence levels, respectively.

The results of the GMM fit with five and six GCs are displayed in Fig. 12, and the corresponding associations, for the solution with five GCs, with the BPT and WHAN classification in Fig. 13. For visualization purposes, the solution with six GCs is also shown, but it fragments the SF region into three parts, which is mostly driven by its banana shape rather than by some physical reason. The inclusion of GC5 reveals the presence of the LINERs in our sample. As expected, the group also appears connected to the passive/retired galaxies class in the WHAN diagram. Conversely, the residual analysis depicted in Fig. 14 shows that the inclusion of an extra five and six components increases the level of variance explained, as one should expect from a maximum likelihood estimator for more complex models, but not significantly. Despite the existence of a physical motivation for the use of an extra group, it does not play a major role in explaining the global data variance on this particular feature space. This suggests that a different choice of feature space or the inclusion of an extra dimension (i.e. emission lines) could be desirable to make the between-group divisions clearer. While our method is capable of automatically recovering groups that resemble previous classifications and provides the means to evaluate their uncertainties, it does not exclude the importance of the domain expert knowledge in order to attribute astrophysical meaning to the results.

8 CONCLUSIONS

In this work we develop a data-driven probabilistic approach to classify galaxies, according to their ionization sources, in a three-dimensional space composed of the log [O III]/H β, log [N II]/H α and log EW(H α) emission lines, which represent a joint BPT–WHAN diagram, using public data from SDSS and the SEAGal/STARLIGHT project.

The results from the parametric GMM are combined with cutting-edge cluster validation methods, also known as internal cluster validation techniques: BIC, ICL, entropy, silhouette and residual analysis. This comprehensive study suggests the existence of four different classes of galaxies, which are capable to explain up to 97 per cent of the data variance in both diagrams.

Table 3. Summary statistics for the $\langle Z/Z_{\odot}\rangle_L$, $\langle \log(t/\mathrm{yr})\rangle_L$ and D_n4000 galaxy properties. Shown are the median, IQR and the first and third quartile for the properties of the GMM, BPT and WHAN groups.

Method	$\langle Z/Z_{\odot}\rangle_L$				$\langle \log(t/\mathrm{yr})\rangle_L$				D_n4000			
Classification	Median	IQR	Q1	Q3	Median	IQR	Q1	Q3	Median	IQR	Q1	Q3
GMM												
GC4	0.50	0.20	0.42	0.62	9.83	0.30	9.66	9.96	1.23	0.08	1.20	1.28
GC1	0.58	0.25	0.47	0.72	9.96	0.18	9.86	10.04	1.32	0.10	1.28	1.38
GC3	0.69	0.37	0.53	0.90	10.03	0.13	9.96	10.09	1.46	0.16	1.38	1.54
GC2	0.97	0.44	0.75	1.19	10.09	0.12	10.02	10.14	1.71	0.29	1.56	1.85
BPT												
SF	0.56	0.25	0.46	0.71	9.95	0.21	9.83	10.04	1.31	0.13	1.25	1.38
Composite	0.78	0.42	0.59	1.01	10.04	0.13	9.97	10.10	1.53	0.20	1.43	1.63
AGN	1.01	0.40	0.82	1.22	10.10	0.11	10.04	10.15	1.73	0.27	1.62	1.89
WHAN												
SF	0.55	0.24	0.45	0.69	9.92	0.25	9.79	10.04	1.29	0.11	1.24	1.35
sAGN	0.65	0.34	0.51	0.85	10.03	0.13	9.96	10.09	1.42	0.15	1.35	1.50
wAGN	0.91	0.41	0.72	1.13	10.05	0.12	9.98	10.10	1.60	0.14	1.54	1.68
Retired/passive	1.06	0.37	0.89	1.26	10.12	0.09	10.06	10.15	1.81	0.18	1.71	1.89

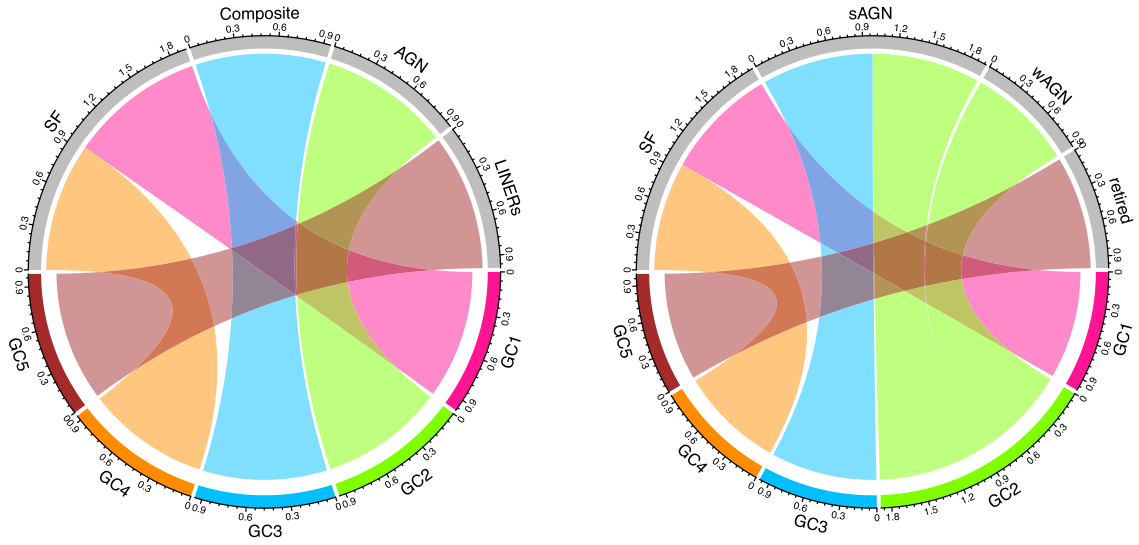

Figure 13. Chord diagrams representing the associations between five GCs and the astronomical classification classes defined on the BPT (with the additional LINER/Seyfert division), left-hand panel, and the WHAN, right-hand panel, diagrams. A thicker connecting ribbon indicates stronger association.

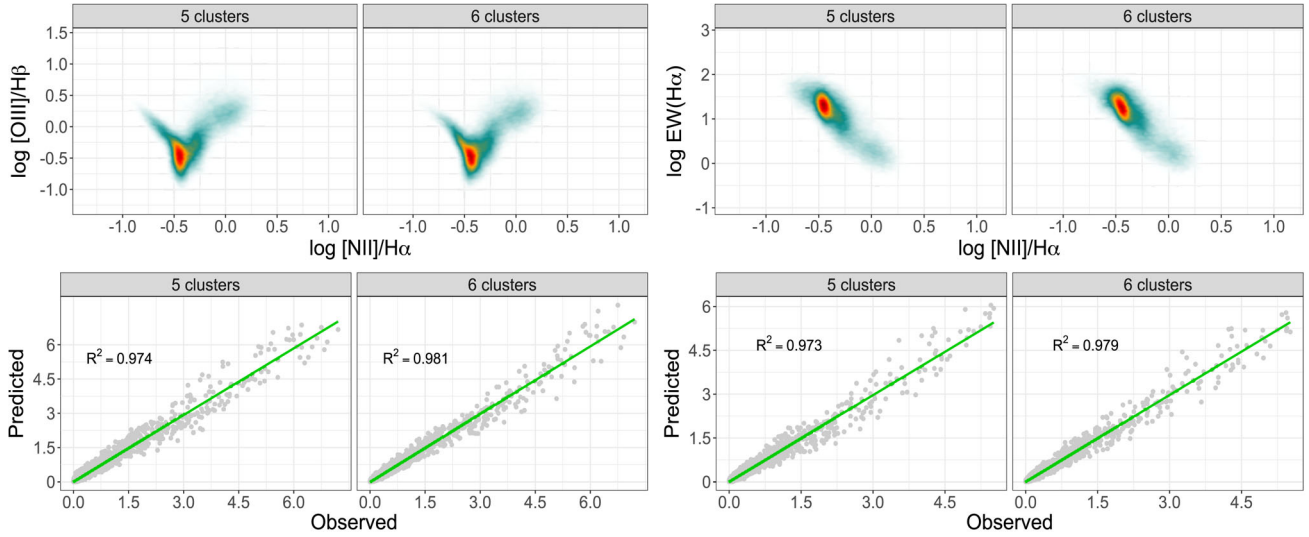

Figure 14. Goodness of fit diagnostics for the five and six GCs projected on to the BPT (left-hand panel) and WHAN (right-hand panel) diagrams. Top: smoothed synthetic data on the BPT and WHAN diagrams as in Fig. 4. Bottom: surface density of the mixture model solution is plotted against surface density of the smoothed observed data. A linear fit of predicted versus observed values is green, and on the left-hand side of each panel we indicate the proportion of variance explained, R^2.

Given the solution with four groups, an external cluster validation approach is employed to compare the GMM results with previous classification schemes based on domain-expert knowledge (i.e. traditional astronomical classes). The results are visualized using a ubiquitous visualization tool in genetics, also known as chord diagram.

Our main scientific results and caveats can be summarized as follows.

(i) The best solution for the GMM, based on maximum likelihood estimation and various quantitative evaluation criteria, has four clusters. The GMM statistically retrieves the existence of the SF, composite and AGN BPT-based groups; and the SF, wAGN, sAGN and retired/passive WHAN-based groups.

(ii) A combination of the GMM results with the external cluster validation technique provides the means to quantify the closeness of each group to their respective counterparts. The SF region (in both diagrams) is divided in two GCs, which might be a consequence of the existence of starburst galaxies populating the top-left wing of the BPT diagram. The composite-BPT region and the sAGN region are both connected to the same GMM-based group, which is mostly due to the lack of a formal composite region in the WHAN diagram. The GMM solution indicates the presence of composite galaxies on the WHAN diagram, in an intermediate region comprising part of the traditional SF and sAGN areas. The wAGN+retired/passive galaxies and the AGN-BPT galaxies are connected to a single GMM group as well, which can be explained by the presence of weak line galaxies that populate the right-wing BPT diagram together with AGN-host galaxies.

(iii) Within the boundaries of the GMM, and in the three-dimensional optical emission-line feature space where we perform the clustering, our data-driven approach does not find a strong

statistical evidence for separate LINER and Seyfert subclasses for the fiducial solution. However, LINERs do emerge as a statistically insignificant subgroup when five GCs are considered.

(iv) To further explore the features of the GMM-based groups against other physical galaxy parameters, not included in the clustering analysis, we compare the statistical distributions of the $\langle Z/Z_{\odot}\rangle_L$, $\langle \log(t/\mathrm{yr})\rangle_L$ and D_n4000 for the GMM, BPT and WHAN classifications. The GMM groups have similar statistical properties compared to the standard diagrams, but with a steeper monotonicity in terms of galactic evolutionary properties.

Our statistical analysis has some limitations and caveats; we address them as follows.

Sample selection. We decided to work with a volume-limited sample to mitigate the Malmquist bias (e.g. Sandage 2000) towards objects with stronger emission lines. If, on the other hand, a magnitude-limited sample were chosen, it would include more fainter/dwarf objects. These objects present larger specific star formation and gas fraction galaxies, and preferentially populate the left-wing region of the BPT diagram. It does not change the overall conclusions regarding the number and location of the GCs. The choice of samples slightly affects the location of the fourth GC responsible for the left-wing region.

Number of clusters. Whilst this fiducial model indicates the presence of four GCs, the results should be informed by astrophysical considerations (e.g. photoionization models). The ICL criterion, which is a regularized version of BIC suggests three groups as optimal case, with the drawback of explaining only up to 90 per cent of the data variance, in contrast to the 97 per cent explained by the use of four GCs. A possible solution to explain the extra variance, while still keeping three groups would be to use a distorted GMM based for instance in a banana shape (see e.g. Laine 2008, for an example of how to sample from a banana-shaped distribution), or to use a non-linear transformation of the feature space (see e.g. Long et al. 2012, as an example of how a non-linear coordinate transformation maps a banana-shaped distributions into a Gaussian one). We should reinforce that the aim here was to build the best model without compromising simplicity and interpretation. Hence, the methodology is a trade-off between predictive power and parsimony, so we prefer to preserve the original space of features and refrained from use non-parametric models (e.g. DBSCAN, k-nearest) or multiparametric distributions as e.g. t-mixture models (Lee & McLachlan 2013). Nonetheless, if physically motivated, a more tailored distribution or feature space should be pursued.

Why Gaussian? We may ask ourselves if a GMM is, in fact, a good approximation to explain the data structure of the BPT–WHAN combined subspaces, specially due to the banana shape of the BPT left wing. One could apply a more flexible non-parametric method, such as DBSCAN, k-nearest neighbours); or to project the data into a non-linear subspace via e.g. kernel principal components analysis (Ishida & de Souza 2013) or isomaps (Wang 2011). However, in any of these options an important feature would be missing – again, simplicity and interpretation; which is a hard compromise to get in general in the machine learning approaches. The GMM may not be the best possible stochastic model to describe the data, but it is a good trade-off between a parsimonious versus an overcomplex model.

Possible follow-ups: (i) the incorporation of physical priors in some based on some flavour of semisupervised technique, in which information regarding the expected number of groups and their locus could be incorporated and refined; (ii) inclusion of additional astrophysical motivated features. For instance, the use of the full width at half-maximum (FWHM) of [O III] could unravel extra groups such as the shock-dominated population, since merges are known to leave imprints in the emission line signal (Leslie et al. 2014); (iii) work directly in the raw spectra using a combination of a manifold and deep learning approaches (Sasdelli et al. 2016) to extract the main spectral features instead of a pre-selected set of emission line; (iv) a comprehensive search for the best lower dimensional subspace able to maximize the discrimination between different galaxy classes.

The analysis herein employed suggests that galaxies with different levels of star formation, with and without supermassive black hole accretion, can be explained by a few classes in low-dimensional spaces. These classes have a measurable mean and standard deviation in the emission-line optical space, and also in the space of other physical parameters, allowing the development of astrophysical models that might be able to predict the physical conditions responsible by the loci occupied by each class.

Summa summarum, this work takes a step forward in the systematic use of machine learning in astronomy. It provides a quantitative and robust recipe for unsupervised astronomical classification and how to combine its output with previous domain knowledge, hence conveying physically interpretable results. Our approach stands out as a valuable tool for future investigations, thanks to its potential to unveil non-trivial relationships in data that may be overlooked by standard procedures. Thus, we strongly advocate for the use of such techniques, especially due to their ability to deal with high-dimensional data sets.

ACKNOWLEDGEMENTS

This work is a product of the 3rd COIN Residence Program (CRP#3). We thank Zsolt Frei for encouraging the accomplishment of this event. CRP#3 was held in Budapest, Hungary, in 2016 August and supported by Eötvös University.

We specially thank E. E. O. Ishida for the organization of the CRP#3 and the fruitful comments and revision of the manuscript. We thank K. Parmar for providing the preliminary external cluster assessment script, and C.-A. Lin for the kind help provided during CRP#3. The authors would also like to thank Professor Plüssmaci for his useful feedback during the entire CRP#3. RSdS thanks Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) process nos 2016/13470-3 and 2012/00800-4 for financial support. MLLD acknowledges Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Instituto de Astronomia, Geofísica e Ciências Atmosféricas da Universidade de São Paulo (IAG-USP) for the financial support. MLLD specially acknowledges S. Rossi and Programa de Excelência Acadêmica (PROEX) for the grant provided for attending CRP#3 and, therefore, accomplishing this work. MLLD also thanks R. Cid Fernandes for providing the treated data set, R. Davies for the help provided and insights in the beginning of this work and P. Coelho for the fruitful discussions and support. MVC-D thanks his scholarship from FAPESP (processes 2014/18632-6 and 2016/05254-9). AK-M thanks the Portuguese agency Fundação para a Ciência e Tecnologia – FCT for financial support (SFRH/BPD/74697/2010). RB was supported through the New National Excellence Program of the Ministry of Human Capacities, Hungary. FG would like to acknowledge the generous support by the Radboud Excellence Initiative.

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

benefited from the following collaborative platforms: OVERLEAF,[9] GITHUB,[10] and SLACK.[11]

In memoriam of Joseph M. Hilbe (30th December 1944–12th March 2017).

REFERENCES

Abazajian K. N. et al., 2009, ApJS, 182, 543
Asari N. V., Cid Fernandes R., Stasińska G., Torres-Papaqui J. P., Mateus A., Sodré L., Schoenell W., Gomes J. M., 2007, MNRAS, 381, 263
Baldwin J. A., Phillips M. M., Terlevich R., 1981, PASP, 93, 5
Baron D., Poznanski D., 2017, MNRAS, 465, 4530
Baudry J.-P., Raftery A. E., Celeux G., Lo K., Gottardo R., 2010, J. Comput. Graphical Stat., 19, 332
Beck R., Dobos L., Yip C.-W., Szalay A. S., Csabai I., 2016, MNRAS, 457, 362
Beck R., Lin C.-A., Ishida E. E. O., Gieseke F., de Souza R. S., Costa-Duarte M. V., Hattab M. W., Krone-Martins A., 2017, MNRAS, 468, 4323
Belfiore F. et al., 2016, MNRAS, 461, 3111
Biernacki C., Celeux G., Govaert G., 2000, IEEE Trans. Pattern Analysis Machine Intelligence, 22, 719
Blanton M. R. et al., 2003a, AJ, 125, 2348
Blanton M. R. et al., 2003b, ApJ, 594, 186
Brinchmann J., Charlot S., White S. D. M., Tremonti C., Kauffmann G., Heckman T., Brinkmann J., 2004, MNRAS, 351, 1151
Bruzual G., Charlot S., 2003, MNRAS, 344, 1000
Cavuoti S. et al., 2015, MNRAS, 452, 3100
Chiosi C., Bertelli G., Bressan A., 1992, ARA&A, 30, 235
Cid Fernandes R., Mateus A., Sodré L., Stasińska G., Gomes J. M., 2005, MNRAS, 358, 363
Cid Fernandes R., Stasińska G., Schlickmann M. S., Mateus A., Vale Asari N., Schoenell W., Sodré L., 2010, MNRAS, 403, 1036
Cid Fernandes R., Stasińska G., Mateus A., Vale Asari N., 2011, MNRAS, 413, 1687
Collister A. A., Lahav O., 2004, PASP, 116, 345
Costa-Duarte M. V., Sodré L., Durret F., 2013, MNRAS, 428, 906
Cover T. M., Thomas J. A., 2006, Elements of Information Theory. Wiley, New York
Cui C., Mao C. X., Zhong J., Zhuang W., 2015, J. Agricultural, Biological, Environmental Stat., 20, 218
Darwin C., 1859, On the Origin of Species by Means of Natural Selection, Or, The Preservation of Favoured Races in the Struggle for Life. J. Murray, London, https://books.google.de/books?id=jTZbAAAAQAAJ
Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, ApJ, 292, 371
Dempster A. P., Laird N. M., Rubin D. B., 1977, J. R. Stat. Soc. Ser. B (Methodological), 39, 1
de Souza R. S., 2015, COINtoolbox.github.io v0.1
De Souza R. S., Ciardi B., 2015, Astron. Comput., 12, 100
De Souza R. S. et al., 2016, MNRAS, 461, 2115
De Vaucouleurs G., 1959, Classification and Morphology of External Galaxies. Springer-Verlag, Berlin, p. 275, http://dx.doi.org/10.1007/978-3-642-45932-0_7
Dressler A., Gunn J. E., 1990, in Kron R. G., ed., ASP Conf. Ser. Vol. 10, Evolution of the Universe of Galaxies. Astron. Soc. Pac., San Francisco, p. 200
Drton M., Plummer M., 2017, J. R. Stat. Soc.: Ser. B (Stat. Methodology), 79, 323
Duda R. O., Stork D. G., Hart P. E., 2001, Pattern Classification. Wiley, New York
Elliott J., de Souza R. S., Krone-Martins A., Cameron E., Ishida E. E. O., Hilbe J., 2015, Astron. Comput., 10, 61
Everitt B., Landau S., Leese M., Stahl D., 2011, Cluster Analysis. Wiley, New York, https://books.google.com.br/books?id=w3bE1kqd-48C
Feigelson E. D., Babu G. J., 2012, Modern Statistical Methods for Astronomy with R Applications. Cambridge Univ. Press, Cambridge
Feltre A., Charlot S., Gutkin J., 2016, MNRAS, 456, 3354
Fisher R. A., 1936, Ann. Eugenics, 7, 179
Fraley C., Raftery A. E., 2002, J. Am. Stat. Assoc., 97, 611
Goto T., 2003, PhD thesis, The University of Tokyo
Gu Z., Gu L., Eils R., Schlesner M., Brors B., 2014, Bioinformatics, 30, 2811
Hastie T., Tibshirani R., Friedman J., 2001, The Elements of Statistical Learning. Springer-Verlag, New York
Hastie T., Tibshirani R., Friedman J., 2009, The Elements of Statistical Learning. Springer-Verlag, New York
Hogan R., Fairbairn M., Seeburn N., 2015, MNRAS, 449, 2040
Ishida E. E. O., de Souza R. S., 2013, MNRAS, 430, 509
Ivezic Z. et al., 2009, BAAS, 41, 366
Juneau S. et al., 2014, ApJ, 788, 88
Karpenka N. V., Feroz F., Hobson M. P., 2013, MNRAS, 429, 1278
Kauffmann G. et al., 2003, MNRAS, 346, 1055
Kewley L. J., Dopita M. A., Sutherland R. S., Heisler C. A., Trevena J., 2001, ApJ, 556, 121
Krone-Martins A., Ishida E. E. O., de Souza R. S., 2014, MNRAS, 443, L34
Kuhn M. A. et al., 2014, ApJ, 787, 107
Kullback S., Leibler R. A., 1951, Ann. Math. Stat., 22, 79
Lacy M. et al., 2004, ApJS, 154, 166
Laine M., 2008, Adaptive MCMC methods with applications in environmental and geophysical models. Finnish Meteorological Institute Contributions, https://books.google.com.br/books?id=JVyDPgAACAAJ
Lamareille F., 2010, A&A, 509, A53
Lamareille F., Mouhcine M., Contini T., Lewis I., Maddox S., 2004, MNRAS, 350, 396
Lee S. X., McLachlan G. J., 2013, J. Stat. Software, 55, 1
Lee K. J., Guillemot L., Yue Y. L., Kramer M., Champion D. J., 2012, MNRAS, 424, 2832
Leslie S. K., Rich J. A., Kewley L. J., Dopita M. A., 2014, MNRAS, 444, 1842
Liddle A. R., 2007, MNRAS, 377, L74
Lindsay B. G., Roeder K., 1992, J. Am. Stat. Assoc., 87, 785
Linnaeus C., 1758, Systema naturae per regna tria naturae :secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis (in Latin)
Lintott C. J. et al., 2008, MNRAS, 389, 1179
Lochner M., McEwen J. D., Peiris H. V., Lahav O., Winter M. K., 2016, ApJS, 225, 31
Long A. W., Wolfe K. C., Mashner M., Chirikjian G. S., 2012, Robotics: Science and Systems. MIT Press, Cambridge, MA, http://dblp.uni-trier.de/db/conf/rss/rss2012.html#LongWMC12
McLachlan G., Krishnan T., 2008, The EM Algorithm and Extensions, 2nd edn. Wiley, Hoboken, NJ, http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+52983362X&sourceid=fbw_bibsonomy
McLachlan G. J., Peel D., 2000, Finite Mixture Models. Wiley, New York, http://opac.inria.fr/record=b1097397
Mateus A., Sodre L., Fernandes R. C., Stasinska G., Schoenell W., Gomes J. M., 2006, MNRAS, 370, 721
Mengersen K., Robert C., Titterington M., 2011, Mixtures: Estimation and Applications. Wiley, New York, https://books.google.com.br/books?id=9JJU_V49MmwC
Morgan W. W., Keenan P. C., 1973, ARA&A, 11, 29
Morgan W. W., Mayall N. U., 1957, PASP, 69, 291
Murphy K. P., 2012, Machine Learning: A Probabilistic Perspective. MIT Press, Cambridge, MA
Pedregosa F. et al., 2011, J. Machine Learning Res., 12, 2825
Poggianti B. M., Barbaro G., 1997, A&A, 325, 1025
R Core Team, 2016, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/

[9] https://www.overleaf.com
[10] https://github.com (de Souza 2015)
[11] https://slack.com/

Richards J. W., Homrighausen D., Freeman P. E., Schafer C. M., Poznanski D., 2012, MNRAS, 419, 1121
Rola C. S., Terlevich E., Terlevich R. J., 1997, MNRAS, 289, 419
Rousseeuw P. J., 1987, Comput. Appl. Math., 20, 53
Sajina A., Lacy M., Scott D., 2005, ApJ, 621, 256
Sandage A., 2000, in Murdin P., ed., Encyclopedia of Astronomy and Astrophysics. IoP Publishing, Bristol, article 1940, doi:10.1888/0333750888/1940
Sasdelli M. et al., 2016, MNRAS, 461, 2044
Schawinski K., Thomas D., Sarzi M., Maraston C., Kaviraj S., Joo S.-J., Yi S. K., Silk J., 2007, MNRAS, 382, 1415
Schawinski K. et al., 2010, ApJ, 711, 284
Schwarz G., 1978, Ann. Stat., 6, 461
Singh R. et al., 2013, A&A, 558, A43
Stasińska G., 2007, preprint (arXiv:0704.0348)
Stasinska G., Fernandes R. C., Mateus A., Sodre L., Asari N. V., 2006, MNRAS, 371, 972
Stasińska G., Costa-Duarte M. V., Vale Asari N., Cid Fernandes R., Sodré L., 2015, MNRAS, 449, 559
Stern D. et al., 2005, ApJ, 631, 163
Stocking G., 1968, Race, Culture, and Evolution: Essays in the History of Anthropology. Univ. Chicago Press, Chicago, https://books.google.de/books?id=aOP43AlmLP8C
Tinsley B. M., 1980, Fundamentals Cosmic Phys., 5, 287
Trouille L., Barger A. J., Tremonti C., 2011, ApJ, 742, 46
Ucci G., Ferrara A., Gallerani S., Pallottini A., 2017, MNRAS, 465, 1144
Vazdekis A., Koleva M., Ricciardelli E., Röck B., Falcón-Barroso J., 2016, MNRAS, 463, 3409
Veilleux S., Osterbrock D. E., 1987, ApJS, 63, 295
Vilalta R., Stepinski T., Achari M., 2007, Data Mining Knowledge Discovery, 14, 1
von der Linden A., Wild V., Kauffmann G., White S. D. M., Weinmann S., 2010, MNRAS, 404, 1231
Wang J., 2011, in Geometric Structure of High-Dimensional Data and Dimensionality Reduction. Springer-Verlag, Berlin, p. 151, doi:10.1007/978-3-642-27497-8_8, http://dx.doi.org/10.1007/978-3-642-27497-8_8
Yan R. et al., 2011, ApJ, 728, 38
York D. G. et al., 2000, AJ, 120, 1579
Zaritsky D., 1993, PASP, 105, 1006
Zhang Y., Zhao Y., 2015, Data Sci. J., 14, 11

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

APPENDIX A: INTERNAL CLUSTER VALIDATION METHODS

A1 Bayesian information criterion

From a Bayesian viewpoint, model selection of a mixture model can be estimated by the integrated likelihood of the model with K components. BIC can be used as a technique that penalizes the likelihood in model selection (Schwarz 1978; Liddle 2007). The higher the value of BIC, the better the result. The BIC for a mixture model log-likelihood is given by

$$\mathrm{BIC}(K) = \log p(x|K, \hat{\theta}_k) - \frac{\nu_K}{2} \log n, \tag{A1}$$

where $\hat{\theta}_k$ is the maximum likelihood estimate of θ_k, and ν_K is the number of free parameters for a model with K components.

A2 Integrated complete likelihood

There is one particular drawback when using BIC to find the best number of clusters: the method works appropriately when each mixture component corresponds to a separate cluster, but this is not always the case. In particular, a cluster may be both cohesive and well distanced from other clusters, without its distribution being Gaussian. Such cluster is best represented with two or more mixture components, rather than a single Gaussian. Hence, the intrinsic number of data clusters may be different from the number of components in the GMM. Biernacki et al. (2000) suggested an alternative to overcome this limitation by directly estimating the number of clusters, as opposed to the number of mixture components; they proposed using the ICL, which can be roughly understood as BIC penalized by mean entropy (Baudry et al. 2010). As a rule of thumb, the number of clusters estimated by ICL is smaller than the number estimated by BIC, due to the additional entropy term. We shall use both indices to constrain lower and upper limits of our solution.

A3 Entropy

A complementary visualization technique to validate the number of clusters based on BIC and ICL is to use the *elbow rule*: a graphical display of entropy variation against the number of clusters. The decrease of entropy at each step serves as a guideline to optimize the number of clusters (Baudry et al. 2010).

A4 Silhouette

The silhouette approach measures the degree of similarity (dissimilarity) of objects within and between clusters (Rousseeuw 1987). It quantifies the common sense that a good clustering algorithm is able to partition the data such that the average distance between objects in the same cluster (i.e. the average intradistance) is significantly lower than the distance between objects in different clusters (i.e. the average interdistance). The technique assigns a value, known as the silhouette width, $s(i)$, to a given cluster solution, which is defined as follows:

$$s(i) = \frac{b(i) - a(i)}{\max a(i), b(i)}, \tag{A2}$$

where $a(i)$ is the average distance between the ith object and all other objects in a given cluster; $b(i)$ is the minimum average distance between the objects in a given cluster and objects in other clusters. Higher silhouette values indicate high-quality clustering solutions.

APPENDIX B: EXTERNAL CLUSTER VALIDATION ALGORITHM

The methodology computes a distance matrix $\mathbf{M}$ where rows correspond to classes and columns correspond to clusters. Each entry M_{ij} captures the (probabilistic) distance between the two data groups. Each class and cluster is modelled as a multivariate Gaussian distribution $f(\boldsymbol{x}) \sim N(\mu, \Sigma)$. From now on, we refer to $f_i(\boldsymbol{x})$ as the Gaussian model for a particular class C_i, and $f_j(\boldsymbol{x})$ as the corresponding Gaussian model for a cluster K_j. We now describe the nature of the metric $\Psi(f_i, f_j)$ used to capture the distance between the two Gaussian models.

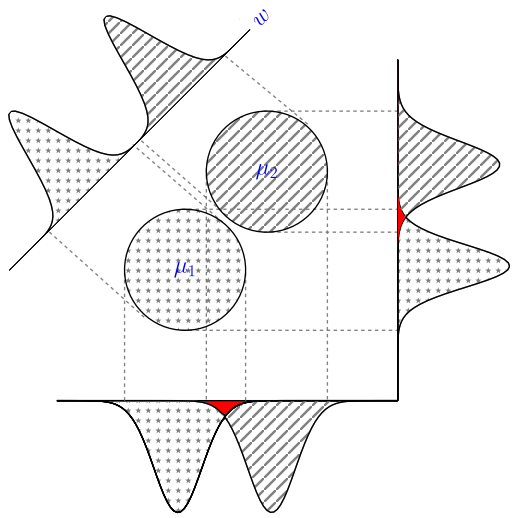

Figure B1. Illustrative figure representing the linear discriminant analysis method. Weight vector $\boldsymbol{w}$, which lies orthogonal to the hyperplane that maximizes the separation between the objects in cluster K_j and class C_i, is used as the dimension over which galaxies are projected.

A straightforward approach to measure the degree of separation $\Psi(f_i, f_j)$ between class C_i and cluster K_j is to use the concept of relative entropy (or KL distance) of two density functions (Cover & Thomas 2006). The relative entropy is the expectation of the logarithm of a likelihood ratio:[12]

$$\Psi(f_i, f_j) = D(f_i||f_j) = \int_x f_i(\boldsymbol{x}) \ln \frac{f_i(\boldsymbol{x})}{f_j(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}. \quad \text{(B1)}$$

This measure can be interpreted as the error generated by assuming that $f_i(\boldsymbol{x})$ can be used to represent $f_j(\boldsymbol{x})$ (or alternatively, the additional amount of information required to describe $f_i(\boldsymbol{x})$ given $f_j(\boldsymbol{x})$). The higher the distance, the higher the dissimilarity between the two distributions.

The metric defined above can be approximated using numerical methods, but the computational cost can become very expensive; integrating over high-dimensional spaces soon turns intractable even for moderately low number of attributes. To address this problem, one final step is necessary. The data are projected into a single dimension $\boldsymbol{w}$, in order to compute the distance function Ψ along that dimension alone. In particular, the proposed solution consists of projecting data objects over a single dimension that is orthogonal to Fisher's linear discriminant (Fisher 1936; Duda, Stork & Hart 2001).[13] The general idea is to find a hyperplane that discriminates data objects in cluster K_j from data objects in class C_i. The weight vector $\boldsymbol{w}$ that lies orthogonal to the hyperplane will be used as the dimension upon which the data objects will be projected. The rationale behind this method is that among all possible dimensions over which that data can be projected, classical linear discriminant analysis identifies the vector $\boldsymbol{w}$ with an orientation that results in a maximum (linear) separation between data objects in K_j and C_i; the distribution of data objects over w provide a better indication of the true overlap between K_j and C_i in multiple dimensions, compared to the resulting distributions obtained by projecting data objects over the attribute axes. Fig. B1 shows our methodology. Weight vector $\boldsymbol{w}$, which lies orthogonal to the hyperplane that maximizes the separation between the objects in cluster K_j and class C_i, is used as the dimension over which data objects are projected.

To add more detail, Fisher's linear discriminant finds the vector $\boldsymbol{w}$ that maximizes the following criterion function: $J(\boldsymbol{w}) = \frac{\boldsymbol{w}^{\mathrm{t}}\mathbf{S}_{\mathrm{B}}\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{t}}\mathbf{S}_{\mathrm{W}}\boldsymbol{w}}$. $\mathbf{S}_{\mathrm{B}}$ is the between-class scatter matrix, defined as the outer product of two vectors: $\mathbf{S}_{\mathrm{B}}(\mu_j - \mu_i)^{\mathrm{t}}(\mu_j - \mu_i)$, where μ_j and μ_i are the mean vectors of $f_j(\boldsymbol{x})$ and $f_i(\boldsymbol{x})$, respectively. $\mathbf{S}_{\mathrm{W}}$ is the within-class scatter matrix, defined as the scatter matrix over the two distributions: $\mathbf{S}_{\mathrm{W}} = \sum(\boldsymbol{x} - \mu_j)^{\mathrm{t}}(\boldsymbol{x} - \mu_j)^{\mathrm{t}} + \sum(\boldsymbol{x} - \mu_i)^{\mathrm{t}}(\boldsymbol{x} - \mu_i)$. It can be shown that a solution maximizing $J(\boldsymbol{w})$ is in fact independent of $\mathbf{S}_{\mathrm{B}}$: $\boldsymbol{w} = \mathbf{S}_{\mathrm{W}}^{-1}(\mu_j - \mu_i)$. Geometrically the goal is to find a vector $\boldsymbol{w}$ so that the difference of the projected means over $\boldsymbol{w}$ is large compared to the standard deviations around each mean (Fig. B1).

[12] The original definition contains $\log_2$, instead of ln; we prefer the latter because it simplifies when the functions are Gaussians; we switch then from a measurement in bits to one in nats.

[13] The use of Fisher's LDA is appropriate here because both LDA and GMMs assume multivariate normality in the components.

This paper has been typeset from a TEX/LATEX file prepared by the author.