Unsupervised Classification of Galaxies. I. ICA feature selection

Tanuka Chattopadhyay, Didier Fraix-Burnet, Saptarshi Mondal

Draft version December 17, 2018
Typeset using LATEX default style in AASTeX62

Unsupervised Classification of Galaxies. I. ICA feature selection

T. Chattopadhyay[1]

D. Fraix-Burnet[2]

S. Mondal[3]

[1] *Department of Applied Mathematics, University of Calcutta, Kolkata, India*
[2] *Univ. Grenoble Alpes, CNRS, IPAG, 38000 Grenoble, France*
[3] *Department of Statistics, Bethune College, Kolkata, India*

(Accepted Dec 10, 2018)

Submitted to PASP

ABSTRACT

Subjective classification of galaxies can mislead us in the quest of the origin regarding formation and evolution of galaxies since this is necessarily limited to a few features. The human mind is not able to apprehend the complex correlations in a manyfold parameter space, and multivariate analyses are the best tools to understand the differences among various kinds of objects. In this series of papers, an objective classification of 362,923 galaxies from the Value Added Galaxy Catalogue (VAGC) is carried out with the help of two methods of multivariate analysis. First, Independent Component Analysis (ICA) is used to determine a set of derived independent components that are linear combinations of 47 observed features (viz. ionized lines, Lick indices, photometric and morphological properties, star formation rates etc.) of the galaxies. Subsequently, a K-means cluster analysis is applied on the nine independent components to obtain ten distinct and homogeneous groups. In this first paper, we describe the methods and the main results. It appears that the nine Independent Components represent a complete physical description of galaxies (velocity dispersion, ionisation, metallicity, surface brightness and structure). We find that our ten groups can be essentially placed into traditional and empirical classes (from colour-magnitude and emission-line diagnostic diagrams, early- vs late-types) despite the classical corresponding features (colour, line ratios and morphology) being not significantly correlated with the nine Independent Components. More detailed physical interpretation of the groups will be performed in subsequent papers.

Keywords: galaxies: general – galaxies: evolution – methods: statistical

1. INTRODUCTION

Investigating the formation and evolution of galaxies is becoming a complicated process with the increased availability of huge databases as a result of instrumental improvements. A good understanding of the underlying physical processes requires synthetic numerical simulations of the formation and evolution of galaxies in an evolving Universe. According to various studies (e.g. see reviews in Silk & Mamon 2012; Genel et al. 2014), classical formation of galaxies have been proposed to follow five major models: (i) the monolithic collapse model, (ii) the major merger model, (iii) the multiphase dissipational collapse model, (iv) accretion and (v) in situ hierarchical merging. But no model uniquely explains the formation of all galaxies.

Corresponding author: T. Chattopadhyay
tchatappmath@caluniv.ac.in

The historical and still most common approach to the classification of galaxies, is based on physical criteria, such as apparent features (i.e. morphology, emission line properties etc.) or more or less understood processes (starbursts, Active Galactic Nuclei (AGN) etc.). Hubble's subjective classification based on galaxy morphology ignores many significant observable features such as kinematics, chemical composition etc. Physically based classifications also rely on only very few attributes among the numerous ones available: colour, size, environment, star formation rate, emission lines. All these classifications are certainly useful to study a specific property, but they cannot embrace the whole diversity and complexity of processes that shaped the galaxies.

With the advent of multi-wavelength and multivariate databases, the goal of many studies has been to find the properties that best characterize the established classes. One such example is given by the Baldwin, Phillips & Tarlevich (BPT) diagrams (Baldwin et al. 1981; Veilleux & Osterbrock 1987) built from a few ratios of emission lines such as [OIII]/Hα, [OIII]/Hβ, [NII]/Hβ, [SII]/Hα and [OI]/Hα. Delimitation lines have been defined empirically to separate different kinds of ionizing sources (AGNs, Low-Ionization Nuclear Emission-Line Regions (LINERs), Seyfert galaxies, star forming regions). But a multivariate analysis using [OIII]/Hα, [NII]/Hβ and EW(Hα) have shown that the empirical classes are not all statistically supported (de Souza et al. 2017). One probable explanation is that several kinds of ionizing sources can be present in a single galaxy, so that a classification built from a two-feature diagram cannot reasonably translate the real complexity of even a single galaxy.

These classification approaches can be called object-based, to be opposed to data-driven science in which some characteristics are revealed by the accumulation of data and attributes. The so-called bimodality of galaxies (Strateva et al. 2001; Eales et al. 2018) is an example of the latter: the observations of millions of galaxies have revealed that for several properties, galaxies show two distribution peaks that do not easily match pre-established physically motivated classes (e.g. Eales et al. 2018; Evans et al. 2018). By chance, this bimodal structure can be seen with the eye on 2D diagrams.

Extragalactic studies have now entered the statistical era challenging our approaches to understand the physics of galaxies and their evolution. In most cases, there is no simple mathematical relationship between a given physical process and the relevant observable attributes. In addition, these attributes are generally multiple so that classifications based on only one or two features are not adequate to represent the physics of galaxies. Despite this caveat, many studies have embarked into complex statistical techniques (Bayesian tools, Deep Learning) to tackle the avalanche of data that astronomy is about to generate with the goal to obtain an automatic classification (e.g. Brescia et al. 2015). However, these approaches, most often multivariate, use supervised learning techniques, i.e. the algorithm is trained on labelled data. The problem here is that the labels come from pre-established classifications based on very few features, in other words the algorithm is trained to find the results of human subjectivity (e.g. Cavuoti et al. 2014).

In this context, one is tempted to apply statistical unsupervised classification, such as a multivariate partitioning analysis to find homogeneous groups, not focusing on only one aspect of the physics of galaxies, but by exploring the cluster structure of the dataset (Fraix-Burnet et al. 2015). The goal is to use data-driven science to study the formation and evolutionary history of galaxies. One basic tool is Principal Component Analysis (PCA) and it has been used by many authors (e.g. Whitmore 1984; Watanabe et al. 1985; Cabanac et al. 2002; Chattopadhyay & Chattopadhyay 2006; Peth et al. 2015). However this is not a classification tool. It may be useful to represent pre-established classes using some of the Principal Components, but its main interest is a reduction of the dimensionality. When the number of attributes is large, distances (e.g. euclidean) become less discriminant so that cluster structures cannot be detected (this is one aspect of the curse of dimensionality). Feature selection, for instance through an objective dimensionality reduction, is then compulsory. But the use of Principal Components to perform a clustering (unsupervised classification) is not recommended since the components with the largest eigenvalues define the axes of maximum variance in the dataset, and these axes are generally not the most discriminative ones to reveal the cluster structure (Chang 1983).

Regarding unsupervised classification, some attempts have been made by K-means cluster analysis (Ellis et al. 2005; Chattopadhyay & Chattopadhyay 2007; Chattopadhyay et al. 2007, 2008; Chattopadhyay et al. 2009a,b; Sánchez Almeida et al. 2010; Fraix-Burnet et al. 2010, 2012; De et al. 2016). Though sophisticated statistical techniques are being developed steadily, unsupervised learning approaches are not widely used across the astronomical community (Fraix-Burnet et al. 2015; de Souza et al. 2017) probably because of the difficulty to relate the multivariate results with physical models. This issue can however be overcome, for instance by comparing statistically the multivariate results with the outcomes of numerical simulations of galaxy and cosmic evolution (e.g. Fraix-Burnet et al. 2012; Genel et al. 2014).

A partitioning of objects into robust groups is possible when the features are independent. Observationally the information is usually summarized into broad-band fluxes (magnitudes), slopes (colours), medium-band and line fluxes (Lick indexes) etc, because they can be easily measured and can generally be yielded by models and numerical simulations. However, their relationship to intrinsic physical processes and their mutual influences are most often quite complex. As a consequence, these direct observables cannot suffice to compare two galaxies and the multivariate aspect of the physics of galaxies must be included as well. This is exactly the purpose of the Independent Component Analysis (ICA) that we will use in this paper.

In this study, we have taken a large dataset from Sloan Digital Sky Survey (SDSS) data archive including various observables regarding morphology, chemical composition and kinematics and used multivariate statistical techniques to explore and explain the underlying diversities. This work presents several novelties for unsupervised classification in astrophysics:

- a large dataset of galaxies
- a large number of features
- the Independent Component Analysis (ICA)
- the adequation of the method (ICA) to the fact that our data are non-Gaussian

Some other studies have used the SDSS to get a large dataset (e.g. Cavuoti et al. 2014) or use many features (up to 4520!, D'Isanto et al. 2018) but they perform supervised classification which, in the case of continuous values, is equivalent to a regression problem. The ICA technique has already been used in astrophysics, mainly for source separation (Pires et al. 2006; Pike et al. 2017; Martins-Filho et al. 2018; Sheldon & Richards 2018) and dimensionality reduction (Richardson et al. 2016; Sarro et al. 2018) but rarely for unsupervised classification (Mu 2007; Das et al. 2015).

This paper is organized as follows. A brief description of the dataset is given in Section 2. The methods are described in Section 3. The results and discussion are included in Sections 4 and 5, respectively. Finally, conclusions are traced in Section 6.

2. DATASET

The NYU Value-Added Galaxy Catalogue (VAGC Blanton et al. 2005; Padmanabhan et al. 2008; Abazajian et al. 2009) is a cross-matched collection of galaxy catalogues maintained for the study of galaxy formation and evolution[1]. It is based on the Sloan Digital Sky Survey Data Release 7 (SDSS-DR7[2]).

In the raw table, 2,506,754 objects are available. We have selected only galaxies, by disregarding QSOs and stars using the SDSS flag and subsequently by removing some obviously wrong classification (such as entries with aberrant redshifts or magnitudes). Non-galaxy objects probably remains, but they should show up in the multivariate classification. We ended up with 865,333 entries. We have then restricted the sample to $z < 0.2$ to avoid too much shift between the wavelength ranges in which the magnitudes are obtained. Finally, only data with a good signal to noise ratio (median S/N per pixel of the whole spectrum > 10) were kept. This leaves us with 362,923 galaxies.

We had to limit the number of attributes to keep the computation tractable. For this, we eliminated redundant properties, i.e. features that bear the same information (such as the Sersic profile in different bands), and selected only a few photometric bands and colours. We were careful to keep most of the physical information so that it does not impact much our analysis based on dimensionality reduction through the ICA. The final set used in our analysis consists of 49 attributes which cover photometry, spectroscopy, morphology, chemical composition and kinematics. Star formation rates and specific star formation rates are also included but not used in the ICA analysis itself since they are not observable features. All these attributes are described in Table A1 and details are given on the source website[3]. Please note that the equivalent width is negative if the line is in emission, EW(Hα) and EW(NII$_{6584}$) are not corrected for possible blending of these two lines, and lines which are absent in spectra are given the value 0.

[1] http://sdss.physics.nyu.edu/vagc/
[2] http://classic.sdss.org/dr7/
[3] http://wwwmpa.mpa-garching.mpg.de/SDSS/DR7/raw_data.html

3. STATISTICAL ANALYSES

3.1. *Shapiro-Wilk test*

The non-Gaussian nature of the dataset has been explored by the statistical Shapiro-Wilk test (Shapiro & Wilk 1965) in which the test statistics is defined by W $= \sum_{i=1}^{n} a_i x_i^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2$, where n is the number of observations, x_i's are ordered sample values and a_i's are the constants generated from the order statistics of a sample from normal distribution. In this work a multivariate extension (Alva & Estrada 2009) has been used. The p value of the test is 2.17×10^{-13}, which is sufficiently small to confidently reject the null hypothesis. Therefore, the dataset is found to be non-Gaussian in nature.

3.2. *Independent Component Analysis*

We have already mentioned that Principal Component Analysis (PCA) has been applied by many authors (Brosche 1973; Whitmore 1984; Murtagh & Heck 1987, etc) but it is not appropriate for clustering and classification (Chang 1983). Furthermore it is applicable to Gaussian data which is not the present case. On the other hand, Independent Component Analysis (ICA) is also a dimension reduction technique, i.e. it reduces the number of observed features p to a number m (m << p) of new variables, the components, but assumes non-Gaussian data (Pfister et al. 2018; Hyvärinen 1998, 1999a,b). The second important difference with PCA is that in addition of being uncorrelated, the components are also mutually independent (for details regarding comparison between PCA and ICA see Section 3 of Chattopadhyay et al. 2013, and references therein).

Mathematically speaking, let X_1, X_2, X_3, ..., X_p be p random vectors (here p features, p = 47) and n (here 362,923) be the number of observations of each X_i, (i = 1, 2, 3, ..., p).

Let X = AS, where S $= [S_1, S_2, S_3, ..., S_p]'$ is a random vector of hidden components S_i, (i = 1, 2, 3, ..., p) such that S_i's are mutually independent and A is a non singular matrix. Then the objective of ICA is to find S by inverting A, i.e., S $= A^{-1}$X = WX. Since ICA is able to separate independent components (sources) present in a signal, W is called the unmixing matrix (for more details see Comon 1994; Chattopadhyay et al. 2013, and references therein). Independence is obtained by maximizing the non-Gaussianity using negentropy.

Presently there is no good method available for the determination of the optimum number of ICs. In this work, the optimum number of ICs have been chosen by the optimum number of Principal Components (PCs) (Albazzaz & Wang 2004; Chattopadhyay et al. 2013; Eloyan & Ghosh 2013), to find m (m << p) (Chattopadhyay & Chattopadhyay 2007; Babu et al. 2009; Fraix-Burnet et al. 2010; Chattopadhyay et al. 2010, 2013). We have first performed PCA to find the significant number of ICs. In PCA the maximum variation with significantly high eigenvalue (viz. $\lambda \sim 1$) was found to be almost 90% for nine PCs. Hence, we have chosen nine ICs (i.e. here m = 9) for cluster analysis (CA). Here it is worth mentioning that the choice of the ICs is a difficult question (see Kairov et al. 2017) but we have checked that the set of most correlated original features (Section 4.1 and Table 1) is stable through multiple runs of the algorithm. Note that for our dataset PCA takes 132 seconds whereas ICA takes less than 1 minute (Intel(R) Core TM i3-2100 CPU @ 3.10GHz, RAM: 2.00 GB, Windows 10, 64 bit operating system) and we have used the library functions publicly available for ICA and PCA in R.

3.3. *K-means cluster analysis*

K-means cluster analysis (CA) is a multivariate technique for finding coherent groups in a dataset giving information of the underlying structure. In this method:

- Each object must belong to a single cluster.
- Each cluster must contain at least one object.
- All the objects are distributed among K clusters by satisfying the first two properties.

Details of algorithm and applications are found in MacQueen (1967); Chattopadhyay et al. (2009b); Chattopadhyay et al. (2010, 2012, 2013); Das et al. (2015).

The number K of groups is an input to the algorithm. The optimum value of K is found as follows. For a particular choice of K initially we compute a distortion measure d_K, given by, $d_K = (1/p) min_x E[(x_K - c_K)'(x_K - c_K)]$ which is the distance of x_K vector (data point) from the centroid c_K of the corresponding group. Then we compute the value

Table 1. Observed attributes with highest correlation coefficients (in parentheses) for each Independent Component.

IC	Most influencial attributes
ICA1	EW(OII_{3729})(0.83), Lick_Fe5015 (0.53)
ICA2	v_{disp} (0.97), J (-0.7)
ICA3	$\sigma_{\rm balmer}$ (-0.60)
ICA4	EW($OIII_{5007}$) (0.98)
ICA5	Lick_Fe5015 (0.83)
ICA6	Sersic_r90_R(-0.99)
ICA7	σ_forb (-0.60), $\sigma_{\rm balmer}$ (0.42)
ICA8	Sersic_amp_R (-0.99)
ICA9	EW(NII_{6584})(0.79), EW(Hα)(0.75), EW(SII_{6731})(0.72)

of jump $J_K = (d_K^{-p/2} - d_{K-1}^{-p/2})$ where p is the total number of variables (here total number of independent components under consideration i.e. p = 9). The above d_K and J_K values are computed for several choices of K starting from K = 1, 2, 3, etc. The maximum value of the curve obtained by plotting J_K versus K gives the optimum value of K (Sugar & James 2003).

In this study, we have performed a K-means CA with respect to the ICs and have found the optimum number of groups to be K = 10. We name the groups K1 to K10.

4. RESULTS

4.1. *Properties of the ICs*

Performing Independent Component Analysis for the VAGC dataset, we have taken nine significant ICs. Then we have computed the correlation coefficients of each component with the attributes to understand their physical meaning. We list only the most influencial attributes in Table 1.

From Table 1, it is clear that the nine ICs represent five kinds of properties: 1) velocity dispersion (ICA 2, ICA 3, ICA 7), 2) ionisation (ICA 4, ICA 9), 3) metallicity (ICA 1, ICA 5), 4) surface brightness (ICA 8) and 5) structural properties (ICA 6). This shows that the ICA successfully retrieve the main physical ingredients of galaxies, while also taking into account their interactions through the linear combinations of the observed features.

This result is indeed very significant. Usually, a given physical process (star formation, nuclear activity, kinematics, radiation characteristics, etc) is investigated through an obvious observable feature, leading by the way to the traditional classifications mentioned in Section 1. Here, each of the ICs is dominated by some features specific to a particular physical property of galaxies, without any a priori selection. In addition, each IC is not limited to the dominant features, it also includes the other ones with some lower weights thus taking into account the complex interplay between observed attributes.

The absence of some traditional features such as colour, line ratios or morphology among the dominant features is striking and will be discussed in Sections 4.3, 4.4 and 4.5 respectively.

These nine independent components are used instead of the initial 47 attributes for Cluster Analysis, very substantially reducing the dimensionality of the dataset while keeping all the physical information.

4.2. *Properties of the galaxies in the ten groups*

The cluster analysis divided the galaxies into ten groups, K1-K10. The distribution of galaxies within these groups is given in Table 2. It appears that the four groups K4, K5, K3, K8, in decreasing importance, already gather 82% of the objects (52% with K4 and K5 only), the K1 group being very small.

Figure A1 shows the distribution of the ten groups in the nine ICs. IC1 and IC3 do not seem very discriminant, but all the others explain the specificities found by the CA algorithm. Many dimension reduction techniques creates new variables (here the IC components) which are linear or non-linear combinations of the original features. They are a mathematical representation of the data that is not intended to have a physical meaning. This is a mathematical space in which the data appear structured in groups. The physical interpretation should aim at understanding these groups in the physical space, not in the IC space. It should be clear that traditional approaches fail since each group is characterized by a complex multivariate physics. It is certainly simpler to summarize galaxies as blue, star forming,

Table 2. Distribution of galaxies in the ten groups found with K-means.

Group	Number of galaxies	Percentage
K1	1,375	0.38
K2	16,325	4.50
K3	68,050	18.75
K4	109,188	30.08
K5	79,576	21.93
K6	17,336	4.78
K7	7,846	2.16
K8	40,045	11.03
K9	10,552	2.91
K10	12,630	3.48

or spiral, but this is obviously very limited. The solution is thus to have more complex descriptions of galaxy classes, and use models and numerical simulations to interpret the results of multivariate analyses.

The boxplots shown in Figure A2 summarize the statistics for the ten groups of the 47 attributes used for the clustering analysis plus SFR, specSFR, Sersic_r50_R and the redshift. The order and the names of the groups given in Table 2 is arbitrary and has been chosen to smoothen the variation of the boxplots from group K1 to K10 for most observed features. This nomenclature thus bears absolutely no physical or temporal evolutionary relationships between these groups.

It is interesting to note that the dispersions are nearly always relatively small, indicating that the groups found by the cluster analysis are quite homogeneous. This is particularly striking for the biggest groups, K4 and K5. There are often large overlaps, but also clearly separated properties distributions between groups in many instances. For some properties, such as Hα_{abs}, J-K and H-K, there is very little variation from group to group.

The small group K1 stands out in many features, especially in EW(OIII$_{5007}$). This group is however close to the groups K2 and K3 for many properties, such as Ca_K$_{\rm abs}$, Sersic_n_R or Lick_G4300. With K4 they share a low velocity dispersion with a clearcut split with respect to the other groups K5 to K10. This split is also visible for Sersic_n_R and also somehow on the redshift diagram.

The specific SFR shows a remarkably regular decrease from K1 to K10, while D4000$_{\rm n}$ has the opposite behaviour.

Groups K6 and K7 gather galaxies which are far much larger than all the other ones. These two groups are not so different in the other observable features, except may be that K7 has a higher average redshift, pointing to possible measurement errors. It is important to recall that most attributes are determined automatically, so aberrant values may appear. However, these two groups are not very small, making such errors (hopefully) improbable. Also, such large sizes are not uncommon (e.g. Guo et al. 2009). Further analyses of the galaxies of these groups, for instance using images, should be done before any physical interpretation. This is beyond the scope of this first paper, we simply emphasize that the CA analysis is able to separate these objects into distinct groups.

4.3. *Colour-magnitude diagram*

The well known bimodality of galaxies is seen on the colour-magnitude diagrams in Fig. 1 with a crescent shape of the distribution of the whole sample, with the so-called red and blue branches (or sequences).

The groups K5, K8, K9 and K10 clearly belong to the red branch, while K2 to K3 are essentially on the blue one. The group K4 spans both branches, but peaks mainly on the red branch and includes the region in between often called the green valley. Interestingly, the groups K6 and K7 that have the largest galaxies in our sample, are both part of this green valley. The very small group K1 appears peculiar, with a very blue part, and 25% of its galaxies belonging to the red branch.

The correspondence between our groups and the rough usual division in red, green or blue regions of the plot is quite good despite the fact that no colour nor magnitude are involved in the independent components used for the classification. This is very likely due to a certain degree of correlation among some properties of galaxies, which multivariate clustering, by determining more subtle and objective categories, is able to stress.

Some groups extend to several regions, such as the big group K4. However, the bivariate colour-magnitude diagram alone cannot reasonably encompass all the physics of galaxies. In this respect, it is interesting to compare our groups

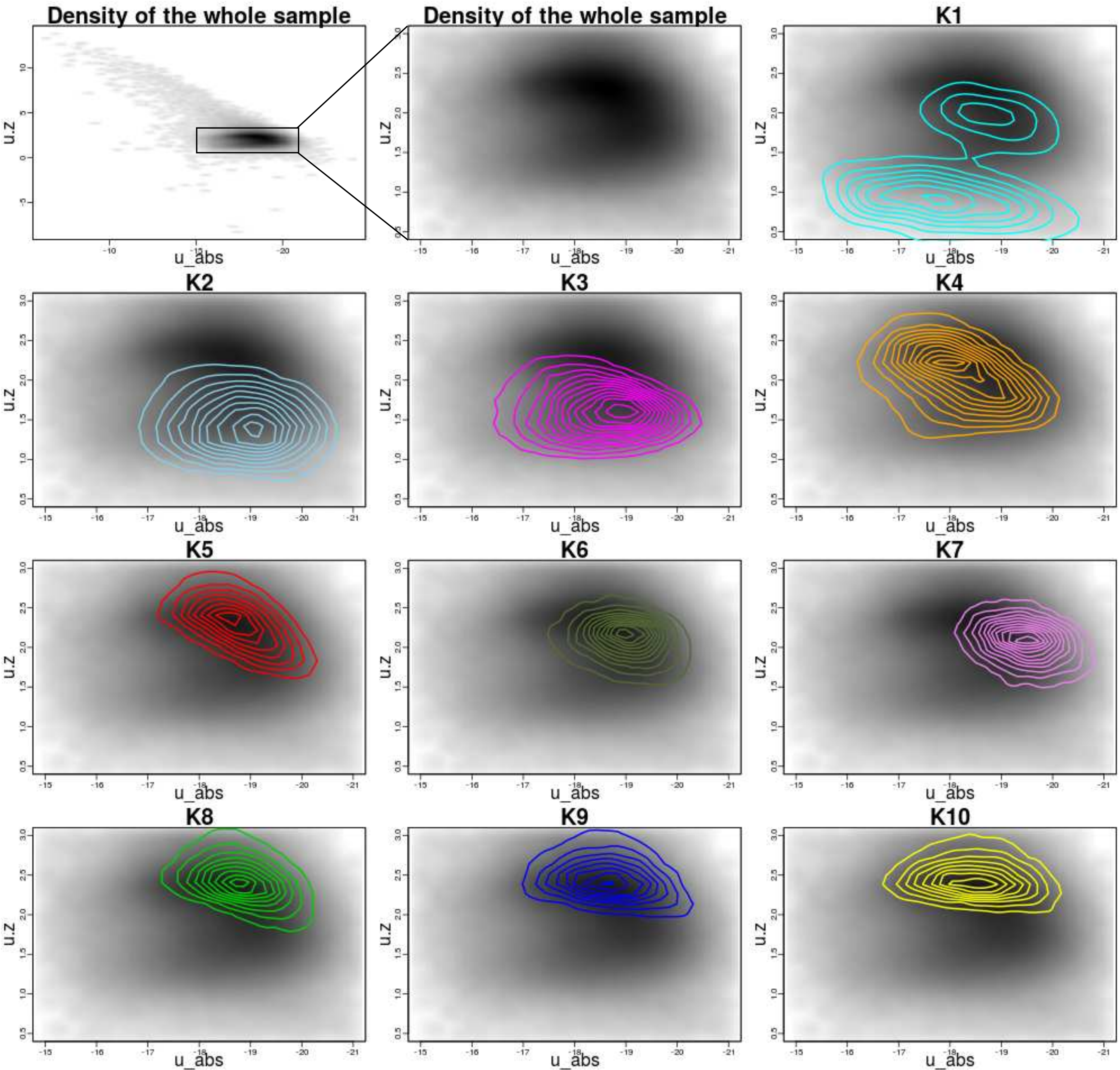

Figure 1. Colour-magnitude (u-z vs. U) diagrams of the whole sample and for each of the groups.

with the Figure 2 of Alatalo et al. (2014) on which some properties, like SFR or EW(Hδ), are plotted. Clearly these properties do not peak and are not confined to only one of the blue, red or green regions, similarly to our groups.

4.4. *Emission-line diagnostic diagrams*

Two emission line ratios were recommended by Baldwin et al. (1981, hereafter BPT) and are often used to discriminate between star forming and AGN-dominated galaxies. This diagram is also the basis for more refined empirical classifications based on theoretical population synthesis and photoionization models (Veilleux & Osterbrock 1987; Kauffmann et al. 2003; Kewley et al. 2006; Kewley et al. 2013). Two other similar diagrams, BPT-SII and BPT-OI, were proposed by Veilleux & Osterbrock (1987). They will be presented in subsequent papers.

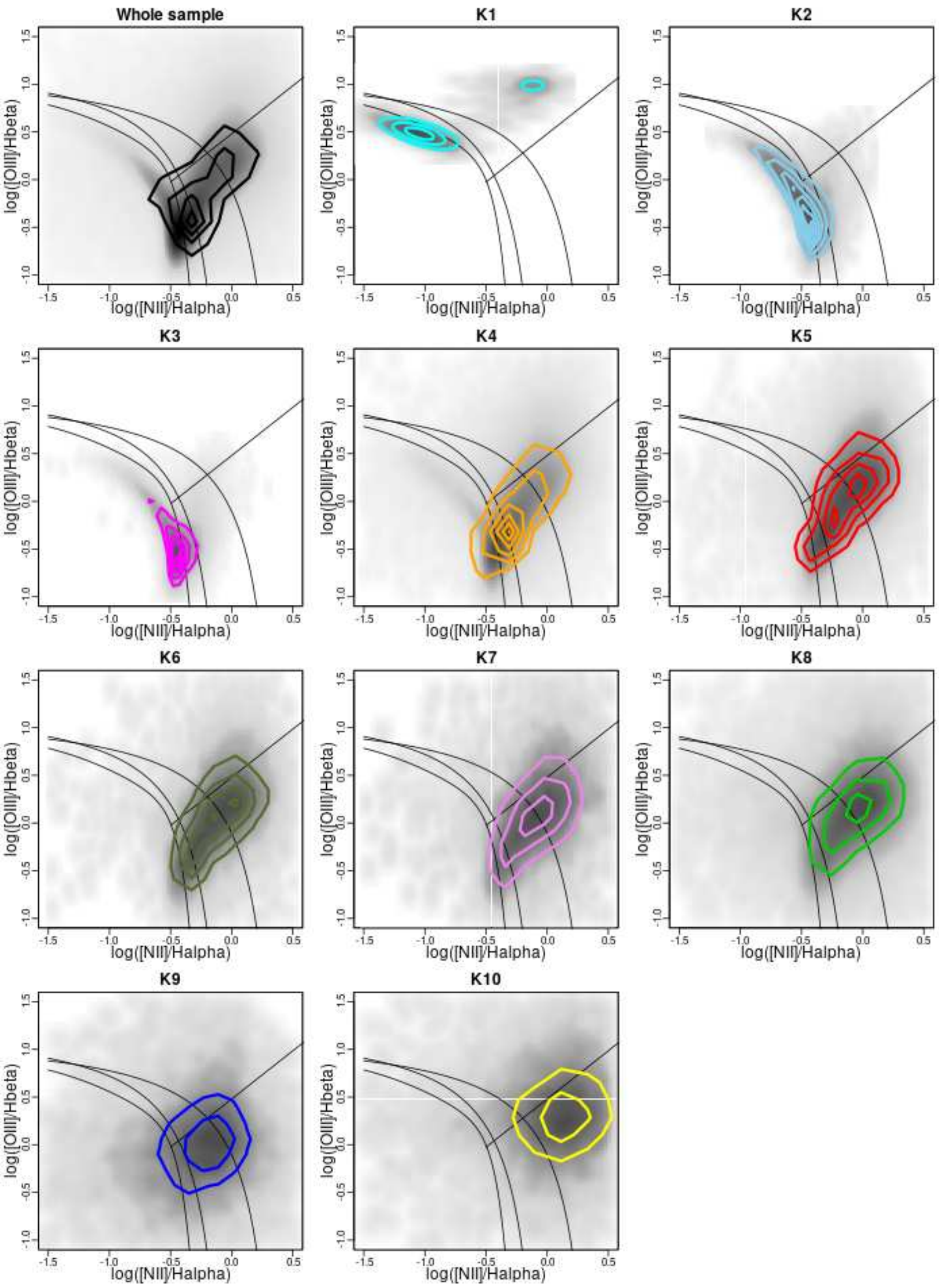

Figure 2. The distribution of galaxies in the groups K1 to K10 in the BPT diagram: [OIII]/Hα vs [NII]/Hβ. Curves are from Kewley et al. (2001); Kauffmann et al. (2003); Stasińska et al. (2006) (right to left respectively). The bottom left zone is for star forming regions, the upper zone is for AGNs and Seyfert galaxies, and the bottom right zone is for LINERs.

The standard classification scheme on this diagram is based on equations of curves separating the different classes. These cuts are sharp, somewhat arbitrary. Its main goal is to distinguish between different ionization source (basically thermal or non-thermal) while the properties used for our clustering analysis are not limited to this peculiar aspect of galaxies. It also appears that the classification based on these diagrams is not clearcut. For instance, "the current LINER classification scheme encompasses two or more types of galaxies, or galaxies at different stages in evolution" (Kewley et al. 2006). Cid Fernandes et al. (2010) proposed a new cut between Seyfert and LINERs to resolve this ambiguous class. They also present an interesting and critical discussion of the classifications based on the BPT diagrams. An important reminder in their discussion is that the star-forming region delimitation is rather arbitrary in the lower part of the diagram where most of the galaxies lie. Also, the upper right part of the diagram is not composed of pure AGNs. Finally, different cuts are proposed by different authors (e.g. Kewley et al. 2001; Lamareille 2010). Interestingly, all these works tend to suggest that more parameters are probably required to fully understand different kinds of galaxies and ionization processes (Richardson et al. 2016).

Indeed, an objective classification with soft frontiers performed with the diagnostics line ratios indicates a somewhat different picture with only four categories, some including for instance sub-divisions like strong or weak AGNs (de Souza et al. 2017). However while being multivariate, this study is only three-dimensional.

An additional limitation of the classifications based on single kinds of ionization processes is that the galaxies are indeed very probably a mixture of different regions (Belfiore et al. 2016).

These limitations on the significance of the cuts should be kept in mind when comparing our multivariate clustering results and diagnostics diagrams (Fig. 2). We see that K2 and K3 have no AGN, hence are pure star forming galaxies, whereas the very small group K1 is quite peculiar with two peaks, one in the pure star forming region, the other one in the pure AGN zone. Groups K4 and K5 are intermediate between star forming galaxies and AGNs. All the other groups (K6, K7, K8, K9 and K10) are LINERs (Fig. 2).

Our result is in the line of the discussion found in the literature as presented above in the sense that our groups can be clearly placed in the different zones of the BPT diagram with some fuzziness. This agreement is remarkable because no line ratios are involved in the ICs used for our classification, only H_α, [OIII] and [NII] but not $H\beta$.

4.5. *Morphology*

In Fig. 3, we plot the density distribution of the concentration index $\mathcal{C}$ = Sersic_r90_R / Sersic_r50_R for all the ten groups. Late-type morphologies are characterized by a low $\mathcal{C}$ ($<$ 2.6, Strateva et al. 2001). Note the strong dichotomy with two peaks at $\mathcal{C}\simeq$ 2.5 and 7.5.

The groups K6 and K7 are nearly entirely made of early-type galaxies while all the other groups are mixtures of both categories, with a higher fraction of late-type galaxies in K2 and K3 and of early-type ones in K5, K10, K8, K9.

The dichotomy between early- and late-type galaxies is thus roughly recovered despite the fact that no morphological indicator such as the concentration index and Sersic_n_R are present in the ICs. This absence probably explains why some of our groups have a mixture of both types, but it is also an indication that the morphology is not very discriminative when the whole multivariate physics of galaxies is considered.

5. DISCUSSION

In this first paper, we intend to focus mainly on the technique used for the multivariate classification of our large sample, and to provide a rather general description of the physical properties of the ten groups we have found to demonstrate that the classification obtained from the IC components is physically meaningful. Naturally, the multivariate grouping seems more complicated than traditional classifications, so that more detailed studies on some specific aspects, such as nucleus activity or SFR, will be made in subsequent papers. For now, from the results described in the previous section, we can distinguish four categories of groups that mainly match usual classes of galaxies.

5.1. *Groups K1, K2, K3*

The three groups K1, K2, K3 belong to the so-called blue sequence and are active sites of star formation as supported by high $H\beta_{abs}$, $H\delta_{abs}$, $H\gamma_{abs}$, specSFR, low metallicity and high EW($OIII_{5007}$).

They are also the less massive objects of the sample, especially K1. They have higher values of $H\delta_{abs}$ thus containing young stellar populations but the widths of its distribution are not similar in these groups. There is maximum scatter in K1 and minimum scatter in K2. They all have a high SFR, but the very small group K1 has a particularly high specSFR. The large scatter in EW($OIII_{5007}$) in the latter group indicates that the galaxies belonging to this K1 group contain gas at high temperature to form new generation of stars. Hence K1 has a stellar population which is a mixture of various ages as a result of star bursts of recent origin.

Though K2 and K3 are dominated by late type galaxies, they also have small fractions of spheroidals. The galaxies in these groups have a low Sersic_n_R lying between 0 and 2 which indicates that the bulges in the spheroidals are not also very pronounced, i.e., these galaxies may be in a formation stage.

5.2. *Groups K4, K5*

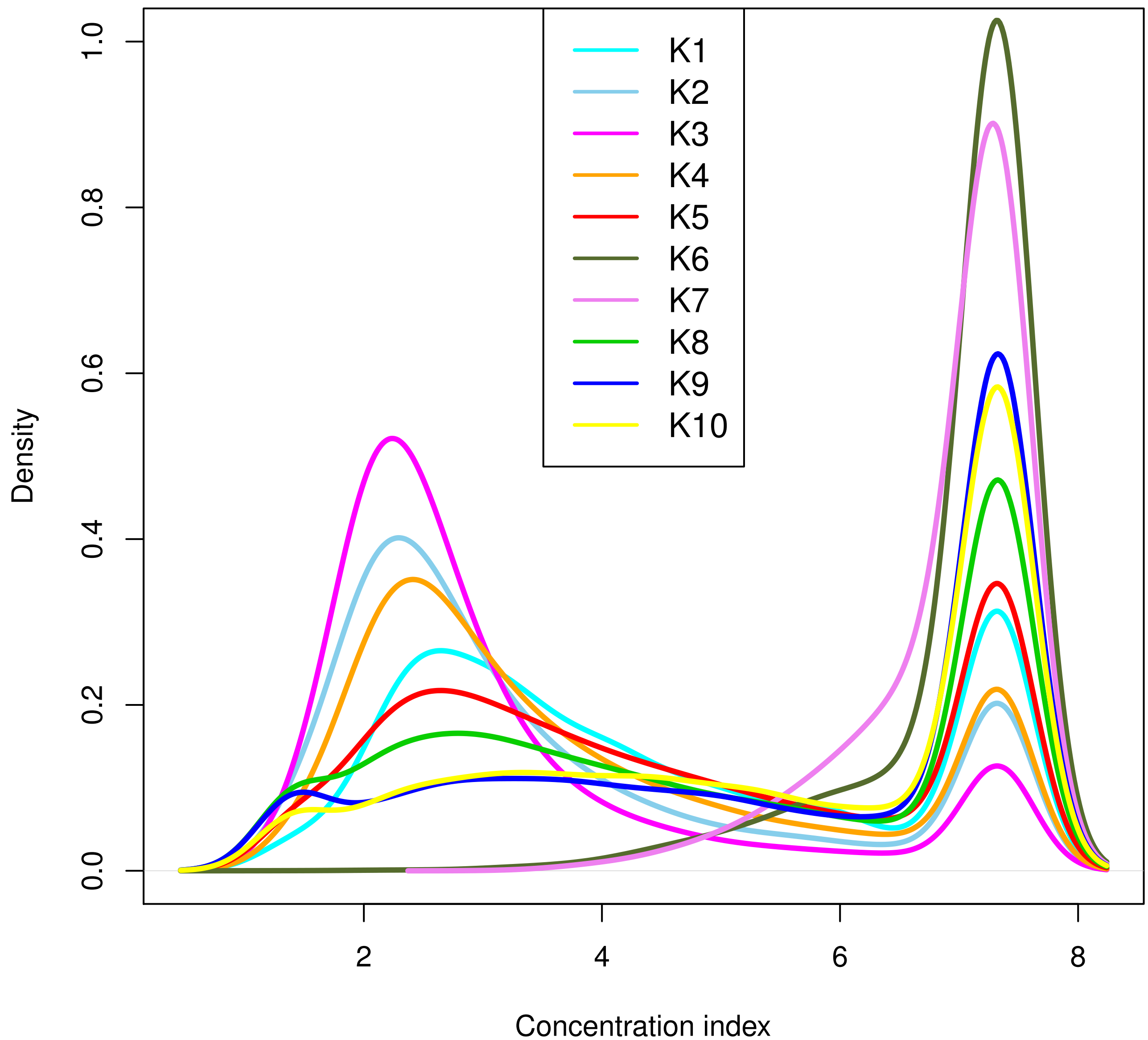

Figure 3. Density distribution of the concentration index $\mathcal{C}$ of galaxies in the groups for K5 to K7.

The two K4 and K5 groups gather about half of our sample. From the colour-magnitude diagram (viz. Fig. 1), it is clear that they are mainly on the red branch with some extensions towards the green and blue regions, especially for K4. However, their Sersic_n_R indexes lie between 2 - 4 which show that some of the galaxies are young and in the formative stages of their bulges and some of them have well developed bulges. The SFR and metallicities in these galaxies are intermediate between highest and lowest values. The two groups differ largely in their light element abundances: K5 has larger value than K4. Regarding Hδ_{abs} the values lie between 0.5 and 2.5 with pronounced scatter. Groups K4 and K5 have globally intermediate values of SFR and more particularly of specSFR with respect to the other groups. Most of the properties like metallicity, SFR, Hδ_{abs}, specSFR, D4000$_{\rm n}$ and $\mathcal{C}$ have large scatters which indicate that K4 and K5 have a certain mixture of old and young populations, as well as of active and passive galaxies with well defined bulges for K5 and more pseudo bulges for K4.

Galaxies in the K5 group have higher velocity dispersion hence are more massive than those belonging the group K4. The group K5 is more abundant in helium enriched population, which is a signature of second generation stars. So K5 has slightly older population than K4. This is also evident from the $D4000_n$ peaks: K5 has a larger peak at higher $D4000_n$ whereas the opposite is true for K4. Also C values show that the two groups have populations of late type spirals, but in much higher proportion in K4.

As a conclusion, despite the fact that K4 and K5 galaxies mainly belong to the red sequence, and would consequently be considered as quenched, they have somewhat average values for most features, and possess some star forming objects.

This result may seem inconsistent, but this is the outcome of a multivariate analysis that gathers galaxies sharing not merely a few properties, but many. However, we cannot exclude that this apparent inconsistency might be due to two possible limitations of our study. The first one is that the clustering algorithm used (K-means) may not be the most suitable technique for the sample, and in particular for the subsamples made with K4 and K5 galaxies. The second explanation might be that even with the information contained in the 47 attributes used in our study and summarized in the 9 ICs used for the classification, it is not possible to distinguish sub-classes within these two big groups. This may be due to several reasons: lack of more distinctive information, uncertainties that smear out differences among variable values of different groups, or the fact that the observables used here are integrated values over an entire galaxy probably mixing up several different regions within these big and complex systems. Further investigations will be made in another paper.

5.3. *Groups K6, K7*

Groups K6 and K7 are globally similar with mostly average values. They are often close to groups K4 and K5, but they have remarkable high C together with very high Sersic_r50_R and Sersic_r90_R. Hence they are made up of very big and large galaxies, and thus spheroidals. They occupy the same region in the colour-magnitude diagram, essentially corresponding to upper part of the so-called green valley. They have a relatively old population as based on $D4000_n$, but with a small fraction of younger stellar populations and some star formation ongoing.

These galaxies thus appear to be in a transitional phase of being quenched but they are really big and massive galaxies which have no equivalent in the red branch members in our sample (red branch that is usually considered as the end fate of all galaxies). The groups K6 and K7 may thus represent the dead end of big galaxies.

5.4. *Groups K8, K9, K10*

The three groups K8, K9 and K10 are rather specific: they have the lowest specSFR, high velocity dispersion, high metallicity, high Sersic indexes and C $>>$ 2.6 for most of their members. All this indicates that they are early type spheroidal galaxies and quenched. They are more concentrated at the center and massive in nature. They have a high abundance of oxygen (viz. Fig. A2, EW($OIII_{4363}$), especially in K10) which might be due to the explosion of massive supernovae (Pop II objects). They have low $H\delta_{abs}$ values with small scatters and high $D4000_n$ values. They have well developed bulges as seen from their high Sersic_n_R values (~ 4, Fig. A2). The forbidden line are pronounced in K10, indicating that these galaxies have enough neutral gas.

5.5. *Independent components*

We have seen that the nine ICs used to perform the cluster analysis represent a complete physical description of galaxies: velocity dispersion, ionisation, metallicity, surface brightness and structure. It is important to realize that they have not been selected on this basis. It should also be clear that using the nine ICs describing five properties is not equivalent to selecting five observables representing these properties, or even nine to match the number of ICs, since the latter are linear combinations of the observable features that reveal some latent multivariate relationships.

The most remarkable outcome is that the groups built from the ICs very satisfactorily match the traditional categorization of galaxies, despite the fact that none of the traditional observables is involved in the nine components. We might see this correspondence from two points of view: either the multivariate analysis is unnecessary, or, as we tend to think, the traditional approach is only an approximation of the complexity of galaxies and is thus incomplete. We obtain more groups than the usual classification, but this is because we are not able to see into a high-dimension space.

Among the observable features involved in the ICs, it is striking to see that the equivalent widths of several lines are important. Also the velocity dispersion and the line widths take their share. The size (radius) and mass (with the

J magnitude) seem to matter. On the contrary, the absence of any colour, of the concentration index, related to the global morphology, or even the index D4000$_n$, is remarkable.

6. SUMMARY AND CONCLUSION

In this study, we have classified a large dataset of galaxies with a large number (47) of morphological, photometric and spectroscopic parameters compiled from VAGC/SDSS data archive. We have used two sophisticated statistical methods. Firstly we have performed a dimension reduction using the ICA analysis which is appropriate for a non-Gaussian dataset such as this one. Secondly, k-means cluster analysis has been used to find structures in the parameter space defined by nine Independent Components to obtain ten coherent groups.

The most correlated observable features involved in the nine ICs represent a complete physical description of galaxies (velocity dispersion, ionisation, metallicity, surface brightness and structure). Despite neither morphology nor colour nor emission-line ratios are among the most influential features in the ICs, our ten groups can be essentially identified as early- or late-type dominated and placed in the colour-magnitude and the line diagnostic (BPT) diagrams in agreement with the literature.

However, our classification is not empirical and is based on a rather complete physical description of the galaxies. The ten groups then allow for a more refined interpretation of galaxy diversity.

Recovering known divisions of galaxies gives us confidence in our unsupervised multivariate clustering analysis, and subsequent papers will be devoted to the astrophysical interpretation and implications of each of the groups we have identified. In particular, the mixture of traditional populations in some of our groups must be understood in the multivariate frame, compared to models and numerical simulations of galaxies that only can provide a multivariate physical understanding, without forgetting the context of evolution.

ACKNOWLEDGEMENTS

We thank Fabrice Lamareille and Emmanuel Davoust for having compiled for us the sample used in this analysis. We warmly thank Malgorzata Siudek for invaluable discussions. TK and DFB are thankful to the Indo-French Centre for Applied Mathematics (Bangalore, India) for providing partial support. We thank the two referees for their detailed and very constructive reports.

Funding for the Sloan Digital Sky Survey (SDSS) has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Aeronautics and Space Administration, the National Science Foundation, the U.S. Department of Energy, the Japanese Monbukagakusho, and the Max Planck Society. The SDSS Web site is http://www.sdss.org/.

The SDSS is managed by the Astrophysical Research Consortium (ARC) for the Participating Institutions. The Participating Institutions are The University of Chicago, Fermilab, the Institute for Advanced Study, the Japan Participation Group, The Johns Hopkins University, Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, University of Pittsburgh, Princeton University, the United States Naval Observatory, and the University of Washington.

REFERENCES

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., et al. 2009, ApJS, 182, 543, doi: 10.1088/0067-0049/182/2/543

Alatalo, K., Cales, S. L., Appleton, P. N., et al. 2014, The Astrophysical Journal, 794, L13, doi: 10.1088/2041-8205/794/1/l13

Albazzaz, H., & Wang, X. Z. 2004, Industrial & engineering chemistry research, 43, 6731, doi: 10.1021/ie049582+

Alva, J. A. V., & Estrada, E. G. 2009, Communications in Statistics - Theory and Methods, 38, 1870, doi: 10.1080/03610920802474465

Babu, G. J., Chattopadhyay, T., Chattopadhyay, A. K., & Mondal, S. 2009, ApJ, 700, 1768, doi: 10.1088/0004-637X/700/2/1768

Baldwin, J. A., Phillips, M. M., & Terlevich, R. 1981, PASP, 93, 5, doi: 10.1086/130766

Belfiore, F., Maiolino, R., Maraston, C., et al. 2016, MNRAS, 461, 3111, doi: 10.1093/mnras/stw1234

Blanton, M. R., Schlegel, D. J., Strauss, M. A., et al. 2005, The Astronomical Journal, 129, 2562, doi: 10.1086/429803

Brescia, M., Cavuoti, S., & Longo, G. 2015, Monthly Notices of the Royal Astronomical Society, 450, 3893, doi: 10.1093/mnras/stv854

Brosche, P. 1973, A&A, 23, 259

Cabanac, R. A., de Lapparent, V., & Hickson, P. 2002, Astronomy and Astrophysics, 389, 1090, doi: 10.1051/0004-6361:20020665

Cavuoti, S., Brescia, M., D'Abrusco, R., Longo, G., & Paolillo, M. 2014, MNRAS, 437, 968, doi: 10.1093/mnras/stt1961

Chang, W.-C. 1983, Applied Statistics, 32, 267, doi: 10.2307/2347949

Chattopadhyay, A. K., Chattopadhyay, T., Davoust, E., Mondal, S., & Sharina, M. 2009a, ApJ, 705, 1533, doi: 10.1088/0004-637X/705/2/1533

Chattopadhyay, A. K., Chattopadhyay, T., De, T., & Mondal, S. 2013, Springer Series in Astrostatistics, Vol. 1, Astrostatistical Challenges for the New Astronomy (Springer-Verlag New York), 185–202

Chattopadhyay, T., Babu, J., Chattopadhyay, A., & Mondal, S. 2009b, Astrophysical Journal, 700, 1768, doi: 10.1088/0004-637X/700/2/1768

Chattopadhyay, T., & Chattopadhyay, A. 2006, The Astronomical Journal, 131, 2452, doi: 10.1086/503160

—. 2007, Astronomy & Astrophysics, 472, 131, doi: 10.1051/0004-6361:20066945

Chattopadhyay, T., Misra, R., Naskar, M., & Chattopadhyay, A. 2007, Astrophysical Journal, 667, 1017, doi: 10.1086/520317

Chattopadhyay, T., Mondal, S., & Chattopadhyay, A. 2008, Astrophysical Journal, 683, 172, doi: 10.1086/589851

Chattopadhyay, T., Sharina, M., Davoust, E., De, T., & Chattopadhyay, A. K. 2012, ApJ, 750, 91, doi: 10.1088/0004-637X/750/2/91

Chattopadhyay, T., Sharina, M., & Karmakar, P. 2010, ApJ, 724, 678, doi: 10.1088/0004-637X/724/1/678

Cid Fernandes, R., Stasińska, G., Schlickmann, M. S., et al. 2010, MNRAS, 403, 1036, doi: 10.1111/j.1365-2966.2009.16185.x

Comon, P. 1994, Signal Processing, 36, 287 , doi: http://dx.doi.org/10.1016/0165-1684(94)90029-9

Das, S., Chattopadhayay, T., & Davoust, E. 2015, PASA, 32, e041, doi: 10.1017/pasa.2015.42

De, T., Fraix-Burnet, D., & Chattopadhyay, A. K. 2016, Communication in Statistics - Theory and Methods, 45, 2638, doi: 10.1080/03610926.2013.848286

de Souza, R. S., Dantas, M. L. L., Costa-Duarte, M. V., et al. 2017, MNRAS, 472, 2808, doi: 10.1093/mnras/stx2156

D'Isanto, A., Cavuoti, S., Gieseke, F., & Polsterer, K. L. 2018, Astronomy & Astrophysics, 616, A97, doi: 10.1051/0004-6361/201833103

Eales, S. A., Baes, M., Bourne, N., et al. 2018, MNRAS, 481, 1183, doi: 10.1093/mnras/sty2220

Ellis, S. C., Driver, S. P., Allen, P. D., Liske, J.b Bland-Hawthorn, J., & De Propris, R. 2005, Monthly Notices of the Royal Astronomical Society, 363, 1257, doi: 10.1111/j.1365-2966.2005.09521.x

Eloyan, A., & Ghosh, S. K. 2013, Computational Statistics & Data Analysis, 58, 383 , doi: https://doi.org/10.1016/j.csda.2012.09.012

Evans, F. A., Parker, L. C., & Roberts, I. D. 2018, MNRAS, 476, 5284, doi: 10.1093/mnras/sty581

Fraix-Burnet, D., Chattopadhyay, T., Chattopadhyay, A. K., Davoust, E., & Thuillard, M. 2012, Astronomy and Astrophysics, 545, A80, doi: 10.1051/0004-6361/201218769

Fraix-Burnet, D., Dugué, M., Chattopadhyay, T., Chattopadhyay, A. K., & Davoust, E. 2010, MNRAS, 407, 2207, doi: 10.1111/j.1365-2966.2010.17097.x

Fraix-Burnet, D., Thuillard, M., & Chattopadhyay, A. K. 2015, Frontiers in Astronomy and Space Sciences, 2, doi: 10.3389/fspas.2015.00003

Genel, S., Vogelsberger, M., Springel, V., et al. 2014, ArXiv e-prints. https://arxiv.org/abs/1405.3749

Guo, Y., McIntosh, D. H., Mo, H. J., et al. 2009, Monthly Notices of the Royal Astronomical Society, 398, 1129, doi: 10.1111/j.1365-2966.2009.15223.x

Hyvärinen, A. 1998, Neurocomputing, 22, 49 , doi: https://doi.org/10.1016/S0925-2312(98)00049-6

—. 1999a, Neural Processing Letters, 10, 1, doi: 10.1023/A:1018647011077

—. 1999b, IEEE Signal Processing Letters, 6, 145, doi: 10.1109/97.763148

Kairov, U., Cantini, L., Greco, A., et al. 2017, BMC Genomics, 18, 712, doi: 10.1186/s12864-017-4112-9

Kauffmann, G., Heckman, T. M., Tremonti, C., et al. 2003, MNRAS, 346, 1055, doi: 10.1111/j.1365-2966.2003.07154.x

Kewley, L. J., Dopita, M. A., Leitherer, C., et al. 2013, ApJ, 774, 100, doi: 10.1088/0004-637X/774/2/100

Kewley, L. J., Dopita, M. A., Sutherland, R. S., Heisler, C. A., & Trevena, J. 2001, ApJ, 556, 121, doi: 10.1086/321545

Kewley, L. J., Groves, B., Kauffmann, G., & Heckman, T. 2006, MNRAS, 372, 961, doi: 10.1111/j.1365-2966.2006.10859.x

Lamareille, F. 2010, A&A, 509, A53, doi: 10.1051/0004-6361/200913168

MacQueen, J. B. 1967, in Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, 281–297. http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html#macqueen

Martins-Filho, W., Griffith, C., Pearson, K., et al. 2018, in American Astronomical Society Meeting Abstracts #231, Vol. 231, 148.26

Mu, B. 2007, PhD thesis, Rochester Institute of Technology

Murtagh, F., & Heck, A., eds. 1987, Astrophysics and Space Science Library, Vol. 131, Multivariate Data Analysis, 236

Padmanabhan, N., Schlegel, D. J., Finkbeiner, D. P., et al. 2008, ApJ, 674, 1217, doi: 10.1086/524677

Peth, M. A., Lotz, J. M., Freeman, P. E., et al. 2015, ArXiv e-prints. https://arxiv.org/abs/1504.01751

Pfister, N., Weichwald, S., Bühlmann, P., & Schölkopf, B. 2018, ArXiv e-prints. https://arxiv.org/abs/1806.01094

Pike, S., Ebisawa, K., Ikeda, S., et al. 2017, ArXiv e-prints, arXiv:1701.05386. https://arxiv.org/abs/1701.05386

Pires, S., Juin, J. B., Yvon, D., et al. 2006, A&A, 455, 741, doi: 10.1051/0004-6361:20053820

Richardson, C. T., Allen, J. T., Baldwin, J. A., et al. 2016, MNRAS, 458, 988, doi: 10.1093/mnras/stw100

Sánchez Almeida, J., Aguerri, J. A. L., Muñoz-Tuñón, C., & de Vicente, A. 2010, ApJ, 714, 487, doi: 10.1088/0004-637X/714/1/487

Sarro, L. M., Ordieres-Meré, J., Bello-García, A., González-Marcos, A., & Solano, E. 2018, MNRAS, 476, 1120, doi: 10.1093/mnras/sty165

Shapiro, S. S., & Wilk, M. B. 1965, Biometrika, 52, 591, doi: 10.1093/biomet/52.3-4.591

Sheldon, K., & Richards, G. 2018, in American Astronomical Society Meeting Abstracts #231, Vol. 231, 250.23

Silk, J., & Mamon, G. A. 2012, Research in Astronomy and Astrophysics, 12, 917, doi: 10.1088/1674-4527/12/8/004

Stasińska, G., Cid Fernandes, R., Mateus, A., Sodré, L., & Asari, N. V. 2006, MNRAS, 371, 972, doi: 10.1111/j.1365-2966.2006.10732.x

Strateva, I., Ivezić, Ž., Knapp, G. R., et al. 2001, AJ, 122, 1861, doi: 10.1086/323301

Sugar, C. A., & James, G. M. 2003, Journal of the American Statistical Association, 98, 750, doi: 10.1198/016214503000000666

Veilleux, S., & Osterbrock, D. E. 1987, ApJS, 63, 295, doi: 10.1086/191166

Watanabe, M., Kodaira, K., & Okamura, S. 1985, The Astrophysical Journal, 292, 72, doi: 10.1086/163133

Whitmore, B. C. 1984, Astrophysical Journal, 278, 61, doi: 10.1086/161768

APPENDIX

A. COMPLEMENTARY TABLES AND FIGURES

Table A1. All the parameters of the dataset used for the analysis.[a]

Parameter	Description
v_{disp}	Estimated velocity dispersion from spectrum
u	u absolute magnitude (log of intensity)
J	J absolute magnitude
Sersic_amp_R	The best fit to the variable "A" in band R (nanomaggies[b]/$arcsec^2$): describes the radial distribution of light
Sersic_n_R	The best fit to the Sersic index "n" in band R
sigma_balmer	Velocity dispersion (σ not FWHM) measured simultaneously in all the Balmer lines in km/s
sigma_forb	Velocity dispersion (σ not FWHM) measured simultaneously in all the forbidden lines in km/s
oii_3729_seqw	The equivalent width of the continuum-subtracted emission line with the other emission lines subtracted off
neiii_3869_seqw	"
oiii_4363_seqw (EW($OIII_{4363}$))	"
oiii_5007_seqw (EW($OIII_{5007}$))	"
hei_5876_seqw	"
oi_6300_seqw	"
h_alpha_seqw (EW(Hα))	"
nii_6584_seqw	"
sii_6731_seqw	"
Hδ_{abs}	Equivalent width for absorption lines
Hγ_{abs}	"
Hβ_{abs}	"
Hα_{abs}	"
Ca_K_{abs}	"
Ca_h_abs	"
Na_d_abs	"
Lick_CN2	Stellar absorption line (Lick) index
Lick_Ca4227	"
Lick_G4300	"
Lick_Fe4383	"
Ca4455	"
Lick_Fe4531	"
Lick_c4668	"
Lick_Hb	"
Lick_Fe5015	"
Lick_Mgb	"
Lick_Fe5270	"
Lick_Fe5335	"
Lick_Fe5406	"
Lick_Fe5709	"
Lick_Fe5782	"
Lick_NaD	"
Lick_Hd_A	"
$D4000_n$	The break in the spectrum at 4000 Å
SFR	Star Formation Rate
specSFR	Specific Star Formation Rate
Sersic_r0_R	The best fit to the variable r_0 in R band (arcsec)
Sersic_r90_R	90% light radius of best fit model in R band (arcsec)
$\mathcal{C}$	Concentration Index: ratio between Sersic_r90_R and Sersic_r50_R
$u - z$	Magnitude u minus magnitude z (u-z)
$J - H$	
$H - K$	

[a] See http://wwwmpa.mpa-garching.mpg.de/SDSS/DR7/SDSS_line.html for more details.

[b] The flux f in nanomaggies is defined in Blanton et al. (2005) as magnitude= $22.5 - 5 \log 10(f)$.

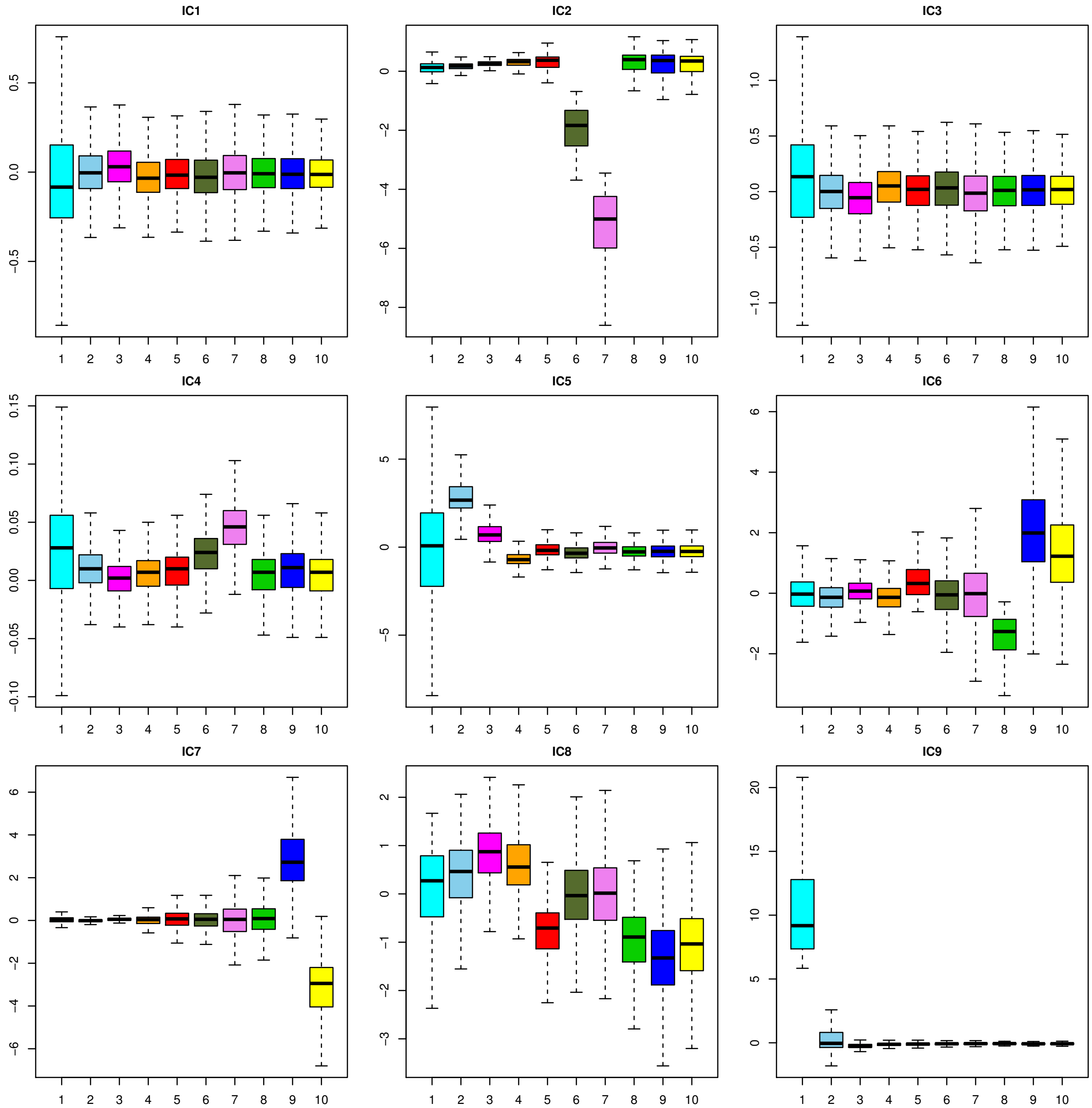

Figure A1. Boxplots for the IC components.

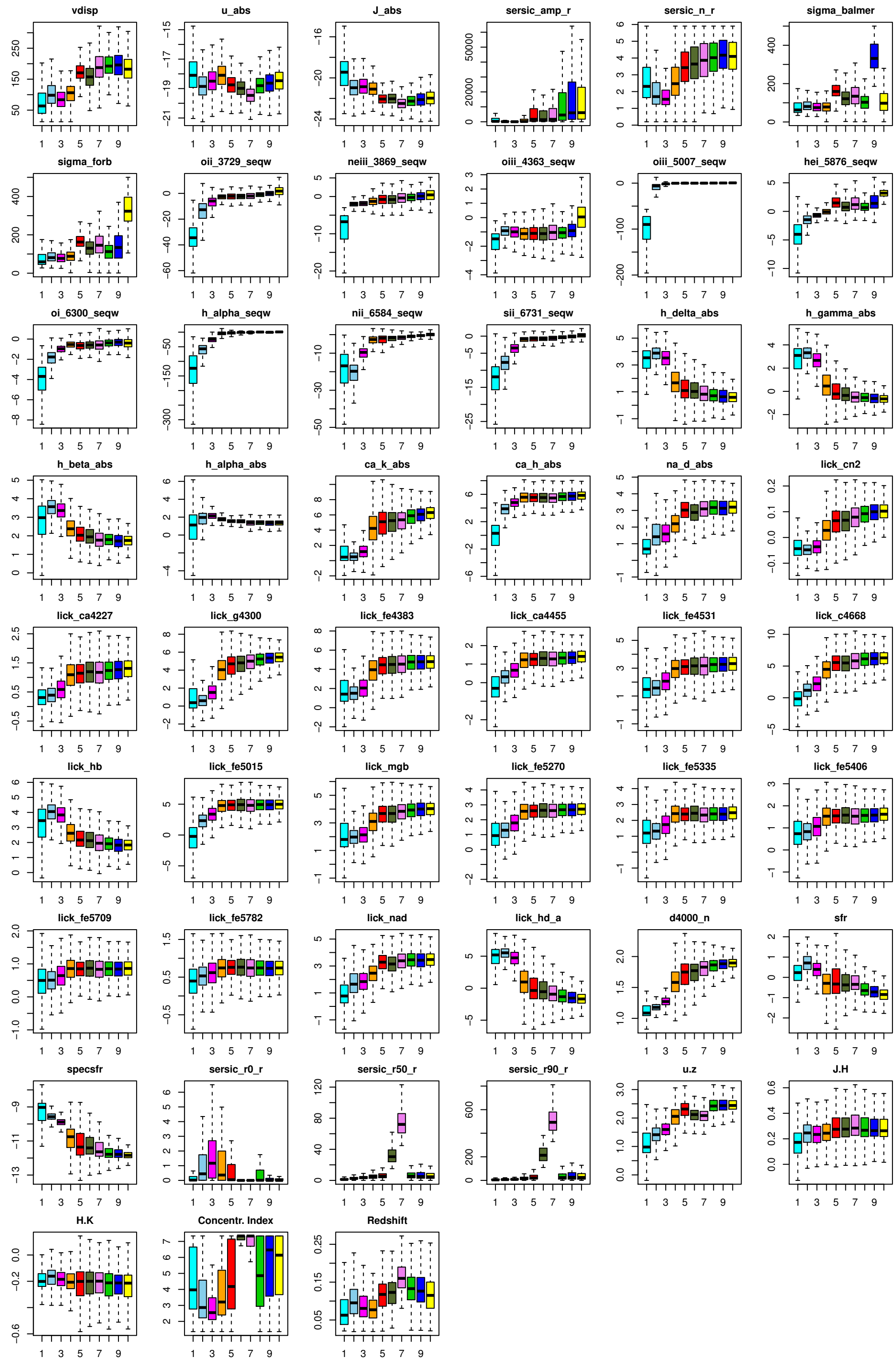

Figure A2. Boxplots for the 47 attributes used for the analysis, plus SFR, specSFR, Sersic_r50_R and the redshift.