Investigation of the effect of bars on the properties of spiral galaxies: a multivariate statistical study

Prasenjit Banerjee\textsuperscript{a}, Tanuka Chattopadhyay\textsuperscript{b}, and Asis Kumar Chattopadhyay\textsuperscript{a}

\textsuperscript{a}Department of Statistics, University of Calcutta, Kolkata, India; \textsuperscript{b}Department of Applied Mathematics, University of Calcutta, Kolkata, India

\textbf{ABSTRACT}
Subjective classification of spiral galaxies is not sufficient for studying the effect of bars on their physical characteristics. In reality the problem is to comprehend the complex correlations in a multivariate parametric space. Multivariate tools are the best ones for understanding this complex correlation. In this work an objective classification of a large set (26,089) of spiral galaxies was compiled as a value added galaxy catalogue from sdss DR 15 virtual data archive. Initially for dimensionality reduction, Independent Component Analysis is performed to determine a set of Independent Components that are linear combinations of 48 observed features (namely ionized lines, Lick indices, photometric and morphological properties). Subsequently a K-means cluster analysis is carried out on the basis of the 14 best chosen Independent Components to obtain 12 distinct homogeneous groups of spiral galaxies. Amongst these, 3 groups are the oldest ones (1.6 Gyr – 5.9 Gyr), while 5 groups fall in the medium aged category (1.4 Gyr – 1.6 Gyr), 2 groups consist of only unbarred spirals, 1 group is the youngest one and the remaining one is an outlier. In many groups there are clear indication of recurrent bar formation phenomena which is consistent with few previous simulation works. In order to study the robustness of the clusters with respect to the method of clustering, a second method of clustering by Gaussian Mixture Modeling Method (GMMBC) is applied.

\section{1. Introduction}
Bars in galaxies are a trace of the dynamical state of the disk (Combes 2009; Athanassoula 2012). They are believed to be temporary structures resulting from gravitational instabilities in the self-gravitating thin disk, and are thought to form in one of two ways: spontaneously, due to internal disturbances within the disk, or through disturbances caused by interactions with neighboring or satellite galaxies. Bars tend to disturb the orbits of stars in the inner regions of galaxies, exciting disk stars into orbits that lie outside the plane of the disk. These stars join the bulge of the galaxy, an example of secular evolution in bulges. Bars also funnel gas and dust into the centers of the galaxies, a process which may trigger bursts of new star formation. Hence, properties of Galaxies are significantly affected by the presence or absence of bars. Effect of bars on the properties of spiral galaxies is an interesting topic as several studies have shown that many important characteristics like star formation activity, color, metallicity etc. may depend on the strength of bar (Athanassoula 1983; Combes and Elmegreen 1993; Martin 1995; Buta and Combes 1996; Ellison et al. 2011; Zhou, Cao, and Wu 2014; Vera, Alonso, and Coldwell 2016). Many other...
works (Weinberg 1985; Debattista and Sellwood 1998; 2000; Athanassoula 2003; Erwin 2019; Garma-Oehmichen et al. 2019; Kim, Choi, and Kim 2020a) show by numerical simulations that bars can effectively transport gas from the outskirt toward the central regions of the barred galaxies. Subsequently this gas undergo interaction with the edges of the bar which produces shock waves. This shocked gas looses angular momentum which accentuates the flow of gas toward the central region and thus produces starburst. Some works show that bars can be destroyed by the presence of large central mass (Roberts, Huntley, and van Albada 1979; Norman, Sellwood, and Hasan 1996; Sellwood and Moore 1999; Athanassoula, Lambert, and Dehnen 2005; Spinoso et al. 2017; Barbuy, Chiappini, and Gerhard 2018; Rosas-Guevara et al. 2019; Guo et al. 2020) This theory indicates that many non-barred disk galaxies might had bars in the past and thus presence and absence of bars are nothing but a recurrent phenomenon of galaxy life (Bournaud and Combes 2002; Berentzen et al. 2004; Gadotti and de Souza 2006; Katz et al. 2018; Pettitt and Wadsley 2018; Hilmi et al. 2020). Many authors have established that inflow of gas is an efficient mechanism for triggering active galactic nuclei (AGN) and they form bulges or pseudo bulges (Kormendy and Kennicutt Jr 2004; Debattista et al. 2005; 2006; Martínez-Valpuesta, Shlosman, and Heller 2006; Aguerri and González-García 2009; Barbuy, Chiappini, and Gerhard 2018; de Lorenzo-Cáceres et al. 2019; Fragkoudi et al. 2020).

In order to investigate the relation between bars and host galaxy colors, different studies inferred that bars are frequently found in late type spiral galaxies those are bluer and less concentrated systems (Barazza, Jogee, and Marinova 2008; Aguerri and González-García 2009). On the contrary other studies found an excess of barred galaxies with redder colors from different samples (Masters et al. 2010; Lintott et al. 2011; Oh, Oh, and Sukyoung 2012; Alonso, Coldwell, and Lambas 2013; 2014; Vera, Alonso, and Coldwell 2016; Cuomo et al. 2019; Kim, Choi, and Kim 2020b).

In connection with star formation activity, many researchers indicate that presence of bars enhance star formation rate (SFR) (Hawarden et al. 1986; Devereux 1987; Hummel et al. 1990), while several others show that bars do not guarantee increase in star formation activity (Pompea and Rieke 1990; Martinet and Friedli 1997; Chapelon, Contini, and Davoust 1999; Kim et al. 2017; Donohoe-Keyes et al. 2019; Newnham et al. 2020; Wang et al. 2020). Similar controversial results are obtained in case of metallicity also (Vila-Costas and Edmunds 1992; Martin and Roy 1994; Ellison et al. 2011; Sánchez-Blázquez et al. 2014).

The above studies are mostly empirical and drawn conclusion either from simulation studies or from a control data sets as a whole, which are prepared in various ways. As the galaxy properties are inter dependent, the corresponding variables constitute a multivariate set up and for data analysis multivariate techniques have to be used. Some authors studied the barred galaxy properties by means of bar properties but that is a subjective classification (Carles et al. 2016; Vera, Alonso, and Coldwell 2016; Kruk et al. 2018; Seo et al. 2019; Cavanagh and Bekki 2020) which is related to algorithm that is trained on label data. The labels arise out of pre-established classifications on very few features, or in other words the algorithm is trained to find results of human subjectivity (e.g., Zooniverse).

In the above context one is tempted to apply statistical (unsupervised) classification e.g., a multivariate partitioning analysis to explore the homogeneous groups of galaxies, not only focusing one or few particular aspects of the physics of galaxies but also by exploring the cluster structure of the data set (Fraix-Burnet, Thuillard, and Chattopadhyay 2015; Chattopadhyay, Fraix-Burnet, and Mondal 2019). One basic tool is Principal Component Analysis (PCA). This is used by various authors (Cabanac, de Lapparent, and Hickson 2002; Chattopadhyay and Chattopadhyay 2006; Whitmore and Forbes 2012; Peth et al. 2016). The main target is the reduction of the dimensionality (i.e., number of variables, here). But use of PC to perform a Clustering

1https://www.zooniverse.org/
(unsupervised) is not recommended since the components with largest eigenvalues are the axes of maximum variance and those are generally not the most discriminative ones to reveal the cluster structure (Chang 1983).

For unsupervised classifications some attempts have been made by K-means cluster analysis on the basis of all the variables (Ellis et al. 2005; Chattopadhyay et al. 2007; Chattopadhyay and Chattopadhyay 2007; Mondal, Chattopadhyay, and Chattopadhyay 2008; Babu et al. 2009; Chattopadhyay et al. 2009; Almeida et al. 2010; Fraix-Burnet et al. 2010; 2012; De, Fraix Burnet, and Chattopadhyay 2016; Modak, Chattopadhyay, and Chattopadhyay 2017; 2020).

Partitioning of objects into robust groups will be more prominent when the features are independent. For this, Independent Component Analysis is used which is applicable to a non-Gaussian data set like the present one for dimensionality reduction. On the basis of the Independent Components (ICs), which are the linear combinations of various observable properties of the spiral galaxies e.g., broad-band line fluxes (magnitudes), slopes (colors), medium band line fluxes (Lick indices), the large data set can be classified (unsupervised) into various homogeneous groups. Then the groups are studied with the help of other estimated properties of the bars, SFR, metallicity, age along with the observed ones.

In this study we have prepared a large data set of spiral galaxies from sdss DR15 and cross-matched with it the corresponding bar properties retrieved from Zooniverse (Galaxy-Zoo). The work contains several novelties for unsupervised classification as follows:

- A large data set of spiral galaxies from sdss DR15, cross matched with zooniverse.
- A large number of observable features.
- A large number of estimated features.
- The application of ICA.
- The justification of ICA to the fact that the present data set is non Gaussian.
- Robustness checking by a widely applicable method like Gaussian Mixture Modeling Method (GMMBC).

Here robustness means that the clusters are robust with respect to the clustering methods, so if we apply different clustering techniques the membership and the overall characteristics remains more or less the same. In most of the previous studies authors retrieved data from sdss and they performed supervised classification. ICA has been used widely for source separation (Pires et al. 2006; Pik et al. 2017; Martins-Filho et al. 2018; Sheldon and Richards 2018) and dimensionality reduction (Richardson et al. 2016; Sarro et al. 2018) but rarely for unsupervised classification (Mu 2007; Das, Chattopadhyay, and Davoust 2015; Modak, Chattopadhyay, and Chattopadhyay 2017; 2020).

The paper is organized as follows: A brief description of the data set is given in Sec. 2. The methods are described in Sec. 3. The results and discussion are included in Sec. 4. Finally, Sec. 5 concludes the summary.

### 2. Data description

The present data is a cross matched collection of galaxy catalogues, used for the study of galaxy formation and evolution. It is based on the sdss Data Release 15 (sdss DR15)\(^2\).

#### 2.1. Data preparataion

The present value added galaxy catalogue of spiral galaxies has been compiled in the following manner:

\(^2\)https://skyserver.sdss.org/dr15/en/tools/search/sql.aspx
(1) The Baldwin, Phillips & Terlevich diagram (Baldwin, Phillips, and Terlevich 1981; Veilleux and Osterbrock 1987; Kauffmann et al. 2003; Kewley et al. 2006; Kewley et al. 2013; Hereafter, BPT diagram). Parameters of the galaxies (viz. Hβ Flux, Hα Flux, Oiii 5007 Flux, Nii 6548 Flux, etc.) are collected by joining the catalogues GalSpecLine on SpecObj through the spectroscopic object id. These parameters of the galaxies are the variables under study.

(2) The Spectral Lick Indices of the galaxies (viz. Lick_Nad, Lick_Ca 4227, Lick_g 4300, Lick_Fe 4383 etc.) are retrieved by joining the catalogues GalSpecIndx on SpecObj through the spectroscopic object id.

(3) The Continuum subtracted emission EW parameters of the galaxies (viz. α-forbidden, Oii 3729 Reqw, Hβ Eqw, Hα Reqw etc.) are collected by joining the catalogues GalSpecLine on SpecObj through the spectroscopic object id.

(4) The Star formation rates and the specific star formation rates of the galaxies are collected by joining the catalogues GalSpecExtra on SpecObj through the spectroscopic object id.

(5) The Photometric properties of the galaxies (viz. u, g, r, i, z, petroR90_r, petroR50_r etc.) are collected from the Galaxy catalogue.

(6) Parameters such as metallicity, logmass, age etc. are collected by joining the catalogues stellarMassStarformingPort on SpecObj through the spectroscopic object id.

(7) The Velocity dispersion the galaxies are collected by joining the catalogues GalSpecInfo on SpecObj through the spectroscopic object id.

(8) The other parameters of the galaxies such as sersic indices, U magnitude etc. are collected from the nsatlas catalogue.

(9) The Spiral properties of the galaxies (viz. p_cw, p_acw, p_edge, p_cs etc.) are obtatined by joining the catalogues zooSpec on SpecObj through the spectroscopic object id.

Now based on the celestial coordinates (RA,DEC) of the galaxies these datasets are cross-matched\(^3\) amongst themselves to obtain our master catalogue. For cross matching we have used the “By position” cross match criteria and the radius is taken as 1 arcsec\(^4\). The cross match is done throughout the sky, not on the basis of any cone or healpix cells\(^5\). Thus the present data set contains only spiral galaxies of different spectroscopic sub classes (viz, UNDEFINED, AGN, AGN_BROADLINE, BROADLINE, STARBURST, STARBURST_BROADLINE, STARFORMING, STARFORMING_BROADLINE etc.). The number of entries are 26,089. Further these master catalogue is crossed with the Hoyle Bar length Catalogue\(^6\), containing 3150 galaxies. After cross matching the number of unbarred spiral galaxies are 24,320 and the number of barred spirals are 1769. In the present data set the barred spirals are denoted by “1”s and the unbarred are denoted by “0”s.

We have limited the number of variables to keep the computation tractable and to reduce noise. Redundant properties such as Sersic profile in different bands have been eliminated since they more or less bear the same information and only a few photometric bands and colors have been selected. Our data set consists of low redshift i.e., \(z \leq 0.06\). We have kept most of the important physical information in our data set, so that it did not greatly impact our analysis based on dimensionality reduction through ICA. Now, our final data set consists of 48 variables which covered spectroscopy, photometry, chemical composition, morphology, and kinematics. All these attributes are described in Table 1 below and details are available on the sdss website\(^7\).

The initial 48 variables in the above table are used for the data analysis and the rest of the variables, mentioned above, have been used for further study (Figure 1).

\(^3\)http://cdsxmatch.u-strasbg.fr/
\(^4\)http://cdsxmatch.u-strasbg.fr/xmatch/doc/CDSXMatchDoc.pdf
\(^5\)http://adass2010.cfa.harvard.edu/ADASS2010/incl/presentations/O01_2.pdf
\(^6\)https://data.galaxyzoo.org/
\(^7\)http://skyserver.sdss.org/dr15/en/help/browser/browser.aspx/#/#/ history=shortdescr+Tables+U
Thus, we are left with a data set containing 26,089 spiral galaxies with 48 different variables for statistical analysis. The data set contains all the variables mentioned over Table 1, except the last eight variables namely Metallicity, z, SFR (M⊙yr⁻¹), log(Mₚ) (Mₚ), Age (Gyr), un(0) barred(1), length_scaled, length_avg. The data set is quite large for computational tractability. Therefore we use the dimensionality reduction (ICA) technique to reduce the size of the data without losing any vital information from it and subsequently we have clustered the data set to observe any coherent groups present in it.

### Table 1. Detail description of the variables of present data set collected for study.

| Variables | Description |
|-----------|-------------|
| Lick_nad  | Stellar absorption line (Lick) index lines |
| Lick_cn2  | h_delta_eqw  |
| Lick_ca4227 | h_gamma_eqw  |
| Lick_g4300 | h_beta_eqw  |
| Lick_fe4383 | h_alpha_eqw  |
| Lick_ca4455 | U  |
| Lick_fe4531 | G  |
| Lick_c4668 | R  |
| Lick_hb | I  |
| Lick_fe5015 | Z  |
| Lick_mgb | J  |
| Lick_fe5270 | H  |
| Lick_fe5335 | K  |
| Dn(4000) | u – g  |
| sigma_balmer | g – r  |
| sigma_forbidden | r – i  |
| oii_3729_reqw | i – z  |
| nii_6584_reqw | u – z  |
| c | Metallicity |

The following variables have been used for further study after finding the homogeneous groups

- Metallicity: Metallicity of best fit template (5 categories: 0.004, 0.01, 0.02, 0.04, or "composite")
- z: Redshift
- SFR (M⊙yr⁻¹): Star-formation rate of best fit
- log(Mₚ) (Mₚ): Best-fit stellar mass of galaxy
- Age (Gyr): Age of best fit
- un(0) barred(1): Indicator variable for barred(1) and unbarred(0) galaxies
- length_scaled: The ratio of bar length to galaxy size
- length_avg: The average bar length scatter per observer, averaged over galaxies being observed

### 3. Statistical methods

Thus, we are left with a data set containing 26,089 spiral galaxies with 48 different variables for statistical analysis. The data set contains all the variables mentioned over Table 1, except the last eight variables namely Metallicity, z, SFR (M⊙yr⁻¹), log(Mₚ) (Mₚ), Age (Gyr), un(0) barred(1), length_scaled, length_avg. The data set is quite large for computational tractability. Therefore we use the dimensionality reduction (ICA) technique to reduce the size of the data without losing any vital information from it and subsequently we have clustered the data set to observe any coherent groups present in it.
3.1. Test for Gaussianity

There are several ways for testing the Gaussianity of a data set, such as Shapiro-Wilk test (Shapiro and Wilk 1965), Anderson Darling Test (Stephens 1974), Kolmogorov-Smirnoff test (Lilliefors 1967) etc. Here we have used the multivariate extension of Shapiro-Wilk Test for checking the Gaussianity of our data set, any other method could have easily opted for. In Shapiro-Wilk Test the test statistic is defined as

$$ W_x = \frac{\left( \sum_{i=1}^{n} a_i x_i \right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, $$

where \( n \) is the number of observations, \( x_i \)'s are the ordered sample values assumed to be present from a standard Gaussian setup under the null hypothesis and \( a_i \)'s are the constants generated from the order statistics of a sample from a normal distribution. As our data set is multivariate in nature, a multivariate extension of the Shapiro-Wilk test (Villasenor Alva and Estrada 2009) have been used. The test statistic for multivariate Shapiro-Wilk test is defined as

$$ W = \frac{1}{p} \sum_{i=1}^{p} W_{z_i}, $$

Figure 1. Images of typical examples of galaxies used in our data set classified as strong barred, weak barred and unbarred galaxies.
where $\textit{p}$ is the number of variables present in our data set, $\textit{Z}_i$'s are the $i^{th}$ vector of $\textit{Z} = \textit{S}^{-1/2}(\textit{X} - \bar{\textit{X}})$, $\bar{\textit{X}}$ and $\textit{S}$ is the sample mean and sample covariance matrix respectively, $W_{Z_i}$’s are the Shapiro-Wilk test statistics for every $\textit{Z}_i$'s. The $p$-value of the test comes out to be $2.2 \times 10^{-16}$, which is quite small. Hence we are more inclined in rejecting the null hypothesis, that our data is from a Gaussian setup i.e., the data set is found out to be Non-Gaussian in nature.

### 3.2. Independent component analysis

PCA has been applied by many authors (Brosche 1973; Murtagh and Heck 2012; Whitmore and Forbes 2012), etc) for several purposes but it is not appropriate for clustering and classification (Chang 1983). Moreover, one of the inherent feature in PCA is that, the data set should be a Gaussian data, but our data set is a non-Gaussian one. ICA is a dimension reduction technique, i.e., it reduces the number of observed variables $\textit{p}$ to a pre-defined number $\textit{m}$ (where, $m \ll p$) of new variables (here the significant IC components). This technique is mainly applicable to non-Gaussian setup (Hyvärinen 1998; 1999a, 1999b; Pfister et al. 2019). Another basic difference between ICA and PCA is that, in PCA the components are assumed to be uncorrelated but not independent where as in case of ICA the components are assumed to be mutually independent amongst each other. For further details regarding the comparison between these two one can consult Sec. 3 of Chattopadhyay, Mondal, and Chattopadhyay 2013a, and references therein.

#### 3.2.1. Method: independent component analysis

let $X_1, X_2, X_3, ..., X_p$ be $p$ random vectors (here, $p = 48$) and $n$ (here, $n = 26,089$) be the number of observations for 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 $S_i$’s, (i = 1, 2, 3, ..., $p$). $A$ is a nonsingular matrix, also known as the mixing matrix. $S_i$’s are mutually independent amongst themselves. The objective of ICA is to find $S$ by inverting $A$, i.e., $S = A^{-1}X = WX$, where, $A^{-1} = W$. $W$ is called the unmixing matrix as it is the inverse of $A$, the mixing matrix. ICA separates the Independent Components (ICs) (sources) present in a mixture (Comon 1994; Chattopadhyay et al. 2013b). To obtain independence, the non-Gaussianity of the data is maximized using negentropy. There are several techniques for ICA such as FastICA, ProDenICA (Hastie and Tibshirani 2003), KernelICA etc. One of them is FastICA algorithm (Hyvärinen and Oja 2000). In this method the ICs are estimated one by one. This algorithm converges very fast and is very reliable. It is the most commonly used algorithm and is also very easy to use.

There is no good method available for the determination of the optimum number of ICs. We generally choose it by using the optimum number of PCs (irrespective of the data being non-Gaussian) (Albazzaz and Wang 2004; Eloyan and Ghosh 2013; Chattopadhyay et al. 2013b), to find $m$ ($m \ll p$) (Chattopadhyay and Chattopadhyay 2007; Babu et al. 2009; Chattopadhyay, Sharina, and Karmakar 2010; Fraix-Burnet et al. 2010; Chattopadhyay et al. 2013b). Another novel criterion of choosing the optimal number of ICs is Maximally Stable Transcriptome Dimension (MSTD) (Kairov et al. 2017).

This technique depends on a fundamental parameter $M$ (effective dimension of the data, as well as the number of ICs computed), whose effects are being investigated on the stability of the ICs. The range of $M$ values are from $M_{\text{min}}$ (here, 2) to $M_{\text{max}}$ (here, 40). For each $M$ ranging from $M_{\text{min}}$ to $M_{\text{max}}$, the data dimension is being reduced to $M$ by PCA and then data has been whitened. Afterwards, in the whitened space the actual signal decomposition is applied by defining $M$ new axes. Each of them maximize the non-Gaussianity of data point projections distribution (Figure 2).
The algorithm for determining the MSTD:

1) Define two numbers $M_{\text{min}}$ and $M_{\text{max}}$ which denote the maximal and minimal possible numbers of computed independent components, respectively.
2) Define a number $T$ (here, $T = 100$). It denotes the number of ICA runs for estimating the components stability.
3) For each $M$ ranging between $M_{\text{min}}$ and $M_{\text{max}}$:
   3.a) Find out the $M$, ICs using the fastICA algorithm and iterate it for $T$ times. Thus we will get a data set of $M/C^2 \times T$, ICs.
   3.b) Now, cluster the newly formed $M/C^2 \times T$, components into $M$ clusters using the agglomerative hierarchical clustering algorithm, where the measure of dissimilarity being $1-\frac{r_{ij}}{1}$. $r_{ij}$ is the Pearson’s correlation coefficient between the components.
   3.c) For each cluster $C_k$ out of $M$ clusters ($C_1, C_2, \ldots, C_M$), the stability index is obtained, using the formula mentioned below:

   $$I_q(C_k) = \frac{1}{|C_k|^2} \sum_{i,j \in C_k} |r_{ij}| - \frac{1}{|C_k|} \sum_{i \neq k} \sum_{j \in C_k} \sum_{i \in C_k} \sum_{j \in C_k} |r_{ij}|$$

   where, $-C_k -$ denotes the size of the $k^{th}$ cluster.
   3.d) The Average Stability Index (ASI) for $M$ clusters is now given by:

   $$S(M) = \frac{1}{M} \sum_{k} I_q(C_k)$$

4) MSTD is given as the point of intersection of the two lines (magenta and the green lines in Figure 3) approximating the distribution of stability profiles. The lines are computed using a simple k-lines clustering algorithm (Agarwal et al. 2005) for $k = 2$. Here MSTD comes out to be 14 (black line in Figure 2).

The clustering quality index as mentioned under 3.c) is used here. It measures the quality of the clustering of ICs after multiple runs with random initial conditions by taking the difference between the average intra-cluster similarity and the average inter-cluster similarity (Himberg, Hyvärinen, and Esposito 2004).

It is further hypothesized that the point of inflection in the distribution of the stability profiles indicates the optimal number of ICs (Figure 2). To find that point, the stability measures are clustered along the two lines (similar to 2-means clustering, but here the lines are taken as centroids instead of points) (Feldman 2003). In this technique, the line (Figure 2, red line) with positive slope grouped the stability profiles with lower values of $M$, while another line (Figure 2, blue line) matched the stability components for the rest. This intersection of these lines (Figure 2,
black line) provided a consistent estimate of the effective number of ICs. This estimate is known as Maximally Stable Transcriptome Dimension (MSTD). This estimate is free of parameters (thresholds) unlike various information theory based criteria (BIC, AIC). It exploits the qualitative change in the character of the stability profile in higher dimensional data.

Three major conclusions can be made from the figures:

1) With the increase of $M$, the average stability of the computed components ($S_{M}^{\text{Total}}$) decreases (Figure 3).

2) $S_{M}^{\text{Total}}$ is characterized by the presence of local maxima, defining certain distinguished values of $M$ that correspond to the (locally) maximally stable set of components (Figure 3).

3) The stability profiles for various values of $M$ can be classified into two, viz. (a) Stability values for which the value of $M$ is low ($M$ upto 9), in this case the stability values are unstable, it shows an irregularity in the trend of the stability profiles over $M$ and (b) the stability values for those forming a large proportion of the components with higher values of $M$, in this case the stability values are stable and gradually decreasing with $M$ (Figure 2).

3.3. Cluster analysis

In our data set we have applied the K-means clustering technique and its robustness in further examined by another clustering technique, viz. Gaussian Mixture Model Based Clustering (GMMBC). In this work we have done clustering on the basis of the 14 Independent Components, which includes spectroscopic, photometric, chemical composition, morphological, and kinematic properties of galaxies.

3.3.1. Optimal choice of clusters

There are several methods for finding the optimal number of clusters, present in the data e.g., technique by Sugar and James (2003), gap statistics (Tibshirani, Walther, and Hastie 2001) and many more . If an inherent clustering is present in the data then it is manifested by any clustering technique. We have used the Dunn index to find out the optimal number of cluster (Dunn 1974) under the method of K-means clustering. Dunn index takes value between 0 to $\infty$. Initially we determined the structures of sub populations (clusters) for varying number of clusters say, $k = 1, 2, 3, 4, \ldots \ldots$ etc. For each such clusters, we have computed the values of the Dunn Index. The value for which the Dunn Index comes out to be highest (here, 0.0063) is taken as the optimal choice of cluster (here, 12) present in our data set and that is the optimal choice of $k$ taken in K-means analysis (Figure 4). This finding is further justified by another technique i.e., GMMBC.
In this technique also, we have initially varied \( k \) over a certain range (say 1, 2, 3, …etc.). As GMMBC is a model based clustering, the choices of \( k \) indicate the choice of the number of models being mixed among themselves and that mixture of models is taken for which the value of BIC is lowest. Here also the optimal number of cluster (i.e., the optimal number of models) appears to be 12 with a BIC value of \(-2.9 \times 10^5\) (Figure 5).

### 3.3.2. Final selection of the reduced data set

Thus we have determined that the data set contains 12 distinct groups amongst themselves and from the previous section, the dimension of the data set can be reduced to 14 from 48 set of variables using MSTD technique (Sec. 3). We have used two different indexes for choosing the best set of 14 ICs from the complete set of 48 ICs namely Average Silhouette Width (ASW) (Rousseeuw 1987) and Within cluster Sum of Squares (WSS). ASW is the given by

\[
W_x = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} S_{ij}}{\sum_{j=1}^{k} n_j},
\]

Where, \( S_{ij} \) is the silhouette index, it is given by

\[
\frac{(b_i - a_i)}{\max(a_i, b_i)}
\]

For every \( j^{th} \) cluster \( C_j \), \( a_i \) is the average distance of object \( i \) from all the other objects in \( C_j \). For any other cluster \( C (\neq C_j) \), \( d(i,C) \) is the average distance of the \( i^{th} \) object in \( C_j \) from all the objects in \( C \). \( b_i \) is the average distance of \( i^{th} \) object in \( C_j \) from all the objects in the nearest cluster \( C \) or in other words, \( b_i = \min_C d(i,C) \). ASW ranges from -1 to 1, a high value indicates a well defined clustering is present in the data set and a low value denotes the data set is homogeneous. ASW and WSS both are used to check whether a clustering is good or poor. WSS is the sum of squares of the distances of the observations \( x \) within a cluster \( C_i \) from the cluster centroid \( r_i \). It is given as:

\[
\text{Figure 4. Dunn Index for different } k.
\]
A good clustering yields a high value of ASW and a small within cluster sum of squares. In this way we choose that particular set of 14 ICs which gives the best clustering with respect to both ASW as well as WSS.

In order to choose 14 ICs at random out of 48 under the above mentioned method we have used systematic sampling. Systematic sampling (Rao et al. 1988) is a type of probability sampling method, in which the samples are drawn from a larger population according to a random starting point but with a fixed and periodic interval. This interval is called the sampling interval. We have used a systematic sampling scheme for the choice of 14 ICs from the set of 48 ICs. We further extend this sampling technique to all possible combinations of the initial random start so that for any random start we have a set of sample in our hand. Moreover we have taken all possible combinations of the sampling intervals also so that for any choice of the sampling interval we have a corresponding set of sample in our hand.

In doing this, some of the samples gets repeated, such as if we start with IC number 1 and goes on taking those ICs with a periodic sampling interval of 4, (with a maximum periodicity of 48, as we have 48 ICs in our hand) then we end up taking IC numbers 1,5,9,13,17,21,25,29,33,37,41,45,1,5 respectively. Here IC number 1 & 5 are repeated in the set, hence this set of ICs are not taken in our analysis. In this way those samples are rejected for our analysis where the ICs gets repeated in choosing the set of 14 different ICs from a sample of 48 ICs.

After getting all the set of ICs we calculate the ASW and WSS (mentioned earlier) for each samples and took that particular sample as our reduced data which gives the maximum ASW or the minimum WSS. ICs starting from number 39 and with a sample interval of 23 (viz., IC39, IC14, IC37, IC12, IC35, IC33, IC8, IC31, IC6, IC29, IC4, IC27, IC2) gives an ASW of 0.188904 & WSS of 5.76606 and satisfies our aforementioned criteria for the selection of the sample. We further rename the set of ICs (IC39, IC14, IC37, IC12, IC35, IC10, IC33, IC8, IC31, IC6, IC29, IC4, IC27, IC2) as (IC1 - IC14) respectively for the sake of our simplicity in addressing the ICs.

Finally, we have applied K-means cluster analysis on this data set by using 14 optimally selected Independent Components as variables and the value of k as 12. Further, a second clustering by GMMBC is performed over the same data set for robustness analysis (Table 2).

The computational time required for performing all the ICA decomposition, the optimal choice of the clusters and the final selection of the reduced data set used in this study is estimated in ~1500 single processor hours using Python interface. As a rough estimate, it takes 6 hours to analyze our dataset with 26089 samples, using an ordinary laptop. In each such analysis, 100 ICA decomposition of different orders have been made. This decomposition of ICs is
Table 2. (The median values of the Age (Gyr), length_avg along with the standard error (in parentheses), and the mean values of the observable variables along with the standard error (in parentheses), for each cluster is given below. The clusters obtained by K-means are denoted by 'K' and the clusters obtained by GMMBC are denoted by 'G'. The mean values for rest of the observable variables along with the standard error (in parentheses) is given in the following URL: https://drive.google.com/file/d/1BnEladJ2A6VYDahR4txNpVbWGxvN8SE/view?usp=sharing).

| Cluster Indices | Sample_Size | Age (Median Values) | Barred_sample_Size | length_avg (Median Values) |
|-----------------|-------------|---------------------|--------------------|---------------------------|
| K1              | 1616        | 2.3(0.05424)        | 301                | 8.87097(0.21045)          |
| K2              | 5257        | 1.609(0.03186)      | 399                | 9.2719(0.17604)           |
| K3              | 4103        | 1.609(0.03445)      | 143                | 8.05755(0.26768)          |
| K4              | 126         | 5.875(0.3058)       | 15                 | 10.68616(0.8906)          |
| K5              | 425         | 1.0152(0.11685)     | 12                 | 7.07733(0.63335)          |
| K6              | 17          | 0.2273(1.55877)     | 1                  | 10.34275(0)               |
| K7              | 23          | 0.7187(0.80196)     | 1                  | 8.10091(0)                |
| K8              | 3723        | 1.434(0.03455)      | 150                | 7.40862(0.26954)          |
| K9              | 1           | 12.25(0)            | 0                  | 10.1304(0.16852)          |
| K10             | 185         | 1.434(0.14679)      | 4                  | 5.63636(0.40655)          |
| K11             | 6356        | 2.1(0.03878)        | 570                | 10.1304(0.16852)          |
| K12             | 4248        | 1.434(0.03203)      | 173                | 8.70386(0.27095)          |

| Cluster Indices | Sample_Size | Age (Median Values) | Barred_sample_Size | length_avg (Median Values) |
|-----------------|-------------|---------------------|--------------------|---------------------------|
| G1              | 432         | 1.434(0.08072)      | 8                  | 7.06626(0.97356)          |
| G2              | 3088        | 1.8(0.03348)        | 363                | 8.77866(0.20262)          |
| G3              | 6803        | 1.2781(0.02093)     | 178                | 6.7802(0.21122)           |
| G4              | 10697       | 1.9(0.02655)        | 816                | 9.75887(0.127)            |
| G5              | 101         | 2.75(0.5013)        | 7                  | 10.0325(1.11173)          |
| G6              | 759         | 1.8(0.12186)        | 13                 | 10.61704(1.21821)         |
| G7              | 2004        | 1.434(0.04803)      | 43                 | 8.27783(0.37798)          |
| G8              | 704         | 1.0152(0.07775)     | 23                 | 7.33386(0.60604)          |
| G9              | 811         | 2.4(0.10432)        | 190                | 10.0183(0.28531)          |
| G10             | 416         | 5.5(0.17722)        | 76                 | 10.2697(0.46843)          |
| G11             | 270         | 1.68(0.17767)       | 51                 | 8.14877(0.5025)           |
| G12             | 4           | 10(2.61632)         | 1                  | 11.65327(0)               |

| Cluster Indices | Sample_Size | log(M_*) | D_*(4000) | U | G | R | I |
|-----------------|-------------|----------|-----------|---|---|---|---|
| K1              | 1616        | 9.90439(0.00835) | 1.2544(0.00299) | -19.43732(0.01757) | -20.70913(0.01635) | -21.3206(0.01633) | -21.67598(0.01666) |
| K2              | 5257        | 9.86208(0.00661) | 1.4662(0.00287) | -18.8964(0.01034) | -20.4194(0.01188) | -21.0993(0.01197) | -21.47548(0.01205) |
| K3              | 4103        | 9.7275(0.00699)  | 1.42718(0.00278) | -18.69309(0.01346) | -20.17005(0.0125) | -20.8404(0.01263) | -21.21499(0.01267) |
| K4              | 126         | 10.72587(0.03902) | 1.90489(0.01165) | -19.1917(0.01796) | -21.11759(0.01708) | -21.98127(0.01763) | -22.42091(0.02761) |
| K5              | 425         | 9.1308(0.02185)  | 1.1808(0.00562)  | -18.6176(0.03978)  | -19.7259(0.0408)  | -20.1414(0.04247) | -20.40716(0.0445)  |
| K6              | 17          | 7.8329(0.37162)   | 1.4090(0.04798)  | -17.14303(0.06172) | -17.9457(0.73786) | -21.02199(0.33941) | -21.18044(0.33999) |
| K7              | 23          | 8.7982(0.06862)   | 1.0491(0.00631)  | -18.3315(0.14112)  | -19.2836(0.12926) | -19.6025(0.11795) | -19.78258(0.12601) |
| K8              | 3723        | 9.62969(0.0748)   | 1.39598(0.00307) | -18.64067(0.03137) | -20.06829(0.01294) | -20.7063(0.01308) | -21.06136(0.01333) |
| K9              | 1           | 8.01(0)           | 1.28818(0)       | -19.7978(0)        | -17.25025(0)      | -17.8538(0)        | -18.1867(0)        |
| K10             | 185         | 9.46335(0.03027)  | 1.35396(0.01373) | -18.30381(0.06301) | -19.70996(0.05604) | -20.33062(0.05465) | -20.67803(0.05457) |
| K11             | 6356        | 10.1918(0.00699)  | 1.57024(0.003)   | -19.15221(0.01021) | -20.76579(0.01048) | -21.527(0.011)    | -21.93120(0.01143) |
| K12             | 4248        | 9.72668(0.00698)  | 1.42778(0.00278) | -18.73717(0.01316) | -20.2036(0.01229) | -20.86652(0.01239) | -21.23347(0.01264) |
Continued.

| Cluster Indices | Sample_Size | $\log(M)$ | $\log(Dn)$ | $\log(U)$ | $\log(G)$ | $\log(R)$ | $\log(I)$ |
|-----------------|-------------|-----------|------------|-----------|-----------|-----------|-----------|
| G1              | 432         | 9.365(0.01722) | 1.3426(0.00994) | 9.365(0.01722) | 1.3426(0.00994) | 9.365(0.01722) | 1.3426(0.00994) |
| G2              | 10697       | 10.12099(0.00453) | 1.29102(0.00126) | 10.12099(0.00453) | 1.29102(0.00126) | 10.12099(0.00453) | 1.29102(0.00126) |
| G3              | 4000        | 432        | 9.365(0.01722) | 1.3426(0.00994) | 9.365(0.01722) | 1.3426(0.00994) | 9.365(0.01722) | 1.3426(0.00994) |
| G4              | 759         | 3088       | 9.785(0.01722) | 1.3426(0.00994) | 9.785(0.01722) | 1.3426(0.00994) | 9.785(0.01722) | 1.3426(0.00994) |
| G5              | 6803        | 9.594(0.01722) | 1.3426(0.00994) | 9.594(0.01722) | 1.3426(0.00994) | 9.594(0.01722) | 1.3426(0.00994) |
| G6              | 101        | 10967      | 9.594(0.01722) | 1.3426(0.00994) | 9.594(0.01722) | 1.3426(0.00994) | 9.594(0.01722) | 1.3426(0.00994) |
| G7              | 795         | 10697      | 10.12099(0.00453) | 1.29102(0.00126) | 10.12099(0.00453) | 1.29102(0.00126) | 10.12099(0.00453) | 1.29102(0.00126) |
| G8              | 204        | 2004       | 9.635(0.01722) | 1.3426(0.00994) | 9.635(0.01722) | 1.3426(0.00994) | 9.635(0.01722) | 1.3426(0.00994) |
| G9              | 704         | 704        | 9.846(0.01722) | 1.3426(0.00994) | 9.846(0.01722) | 1.3426(0.00994) | 9.846(0.01722) | 1.3426(0.00994) |
| G10             | 811         | 811        | 9.846(0.01722) | 1.3426(0.00994) | 9.846(0.01722) | 1.3426(0.00994) | 9.846(0.01722) | 1.3426(0.00994) |
| G11             | 5257        | 5257       | 9.846(0.01722) | 1.3426(0.00994) | 9.846(0.01722) | 1.3426(0.00994) | 9.846(0.01722) | 1.3426(0.00994) |
| G12             | 4          | 4          | 9.846(0.01722) | 1.3426(0.00994) | 9.846(0.01722) | 1.3426(0.00994) | 9.846(0.01722) | 1.3426(0.00994) |

Table 2. Continued.
| Cluster Indices | Sample Size | G-R      | R-I       | I-Z       | U-Z       | J-H      | H-K      |
|----------------|-------------|----------|-----------|-----------|-----------|----------|----------|
| K1             | 1616        | 0.61146(0.00341) | 0.35539(0.00166) | 0.26177(0.00171) | 2.50043(0.01098) | 0.72359(0.00568) | 0.43645(0.00606) |
| K2             | 5257        | 0.68736(0.00201) | 0.37617(0.00095) | 0.28002(0.00103) | 2.8591(0.00793) | 0.73397(0.00451) | 0.4314(0.00466)  |
| K3             | 4103        | 0.67019(0.00235) | 0.37471(0.00118) | 0.27525(0.00161) | 2.7971(0.00861) | 0.76172(0.00562) | 0.44155(0.00577) |
| K4             | 126         | 0.86367(0.00633) | 0.42165(0.00452) | 0.32938(0.00546) | 3.54056(0.03394) | 0.70069(0.01356) | 0.35712(0.01082) |
| K5             | 425         | 0.41384(0.00715) | 0.26574(0.00404) | 0.17749(0.00355) | 1.96702(0.02333) | 0.77079(0.01829) | 0.51535(0.01996) |
| K6             | 17          | 3.22742(0.67742) | 0.15845(0.18426) | -3.24598(0.93807) | 0.79144(0.79693) | 0.82859(0.09596) | 0.48271(0.08432) |
| K7             | 23          | 0.3189(0.02483)  | 0.18001(0.01508) | 0.13755(0.01397) | 1.58497(0.05968) | 0.77857(0.1127)  | 0.38922(0.06759) |
| K8             | 3723        | 0.63801(0.00252) | 0.35506(0.00128) | 0.26178(0.00131) | 2.68247(0.00883) | 0.76818(0.00632) | 0.4674(0.00648)  |
| K9             | 1           | 0.60358(0)      | 0.33287(0)      | 0.12040(0)      | 8.93096(0)      | 0.187(0)      | 1.08(0)    |
| K10            | 185         | 0.62065(0.01164) | 0.34741(0.00594) | 0.25872(0.00618) | 2.63294(0.03915) | 0.81959(0.03145) | 0.49461(0.03302) |
| K11            | 6365        | 0.76131(0.00188) | 0.4041(0.0019)  | 0.31391(0.00195) | 3.0929(0.00683) | 0.73703(0.00303) | 0.41661(0.00317) |
| K12            | 4248        | 0.66292(0.00235) | 0.36695(0.00159) | 0.27412(0.00165) | 2.77042(0.00818) | 0.76692(0.00562) | 0.43787(0.00577) |

| Cluster Indices | Sample Size | G-R      | R-I       | I-Z       | U-Z       | J-H      | H-K      |
|----------------|-------------|----------|-----------|-----------|-----------|----------|----------|
| G1             | 432         | 0.62095(0.00689) | 0.35303(0.00391) | 0.26257(0.00424) | 2.63833(0.02381) | 0.90199(0.02149) | 0.52295(0.022)  |
| G2             | 3088        | 0.60308(0.00247) | 0.34852(0.00126) | 0.2535(0.00122)  | 2.51236(0.00821) | 0.72658(0.00502) | 0.43753(0.00517) |
| G3             | 6803        | 0.6123(0.00369)  | 0.34915(0.00091) | 0.25396(0.00093) | 2.60427(0.00591) | 0.79493(0.00488) | 0.48347(0.00501) |
| G4             | 10697       | 0.75581(0.00725) | 0.40164(0.00626) | 0.30777(0.00063) | 3.07546(0.00454) | 0.7059(0.00239)  | 0.394(0.00256)  |
| G5             | 101         | 1.14197(0.1478)  | 0.27309(0.11667) | -0.33705(0.23185) | 2.38619(0.21734) | 0.74501(0.03715) | 0.493(0.03616)  |
| G6             | 759         | 0.7627(0.0559)   | 0.42452(0.00922) | 0.31725(0.00322) | 2.3737(0.03104)  | 0.78066(0.01374) | 0.46933(0.01393) |
| G7             | 2004        | 0.67452(0.00332) | 0.37561(0.00178) | 0.28329(0.00189) | 2.82068(0.01205) | 0.8022(0.00945)  | 0.46205(0.00947) |
| G8             | 704         | 0.44103(0.00492) | 0.27228(0.00314) | 0.18547(0.00296) | 2.05821(0.01635) | 0.82398(0.01488) | 0.51661(0.01582) |
| G9             | 811         | 0.74156(0.00487) | 0.39747(0.00772) | 0.31498(0.00776) | 2.95766(0.01799) | 0.73847(0.00631) | 0.44734(0.00622) |
| G10            | 416         | 0.86792(0.00318) | 0.42896(0.00217) | 0.3327(0.0023)   | 3.52479(0.01341) | 0.69446(0.00656) | 0.3598(0.00661)  |
| G11            | 270         | 0.50498(0.01069) | 0.30478(0.00338) | 0.21357(0.00484) | 2.18964(0.03459) | 0.7559(0.01862)  | 0.42569(0.01913) |
| G12            | 4           | 1.07286(0.30623) | 0.45926(0.08625) | 0.41868(0.18051) | 5.20721(1.29849) | 0.568(0.12765)  | 0.34375(0.23419) |
4. Results

In the present work we have compiled a data set up to a red shift of \( z < 0.06 \), consisting of unbarred, weak barred and strong barred spiral galaxies and including starburst, AGN and LINER. At first we have checked the data set for Gaussianity and found it to be non-Gaussian. We performed Independent Component Analysis for dimensionality reduction and used MSTD to find out the optimal number of Independent Components. Then we clustered the data set with respect to these Independent Components by K-means cluster analysis followed by finding the number of optimum groups. The optimum number of ICs is 14 and the number of coherent groups is 12 (Table 2). From Table 2, it is clear that group 9 (K9) is an outlier and groups 6 and 7 (K6 and K7) contain a very small number of galaxies. Hence we have not considered K9. Subsequently we performed Gaussian Mixture Model Based Clustering (GMMBC) for checking the robustness of the groups. We have found similar number of groups with one group containing only 4 members like K9. Remaining 11 groups found by both the methods are more or less compatible with respect to membership as well as average values of the variables (K1 \( \rightarrow \) G2, K2 \( \rightarrow \) G6, K3 \( \rightarrow \) G7, K4 \( \rightarrow \) G10, K5 \( \rightarrow \) G8, K6 \( \rightarrow \) G9, K7 \( \rightarrow \) G11, K8 \( \rightarrow \) G1, K9 \( \rightarrow \) G12, K10 \( \rightarrow \) G3, K11 \( \rightarrow \) G4, K12 \( \rightarrow \) G5). Therefore we have discussed the physical properties of the groups found with respect to K-means cluster analysis.

4.1. Properties of the ICs

In the analysis of ICA, we randomly selected set of ICs for which the variation in the multivariate set up is maximum with respect to Dunn index (Figure 4). For the optimum set there are 14 ICs. ICs are actually linear combinations of several variables (here 48 observable variables) with various co-efficients. Few ICs are denoted by some features specific to a particular physical property of galaxies without any prior selection. At the same time each IC is not limited by the dominant feature. It also includes other with some lower weights and thus takes into account the complex interplay between observable variables. Table 3 shows that among 14 ICs, 7 represent about five kind of properties:

1) Metallicity (IC1,IC2,IC5),
2) Balmer absorption feature and low level ionization (IC3),
3) Color (IC4),
4) Velocity dispersion (IC14),
5) Metallicity and high level ionization (IC6, IC11).

Table 3. (Observed variables with highest correlation coefficients with significant ICs.)

| IC  | Influential observed variables                                      |
|-----|---------------------------------------------------------------------|
| IC1 | Lick_Fe 5406 (0.54)                                                  |
| IC2 | Lick_Ca 4455 (-0.30)                                                 |
| IC3 | Lick_Hb (0.62), EW (NII 6584) (0.65), EW (H\(_{\alpha}\)) (0.57), EW (H\(_{\beta}\)) (0.65), EW (H\(_{\gamma}\)) (0.65), EW (H\(_{\delta}\)) (0.65) |
| IC4 | i-z (0.30)                                                           |
| IC5 | Lick_Nad (0.76), Lick_Ca 4227 (0.70), Lick_g 4300 (0.76), Lick_Fe 4383 (0.74), Lick_Fe 4531 (0.71), Lick_C 4668 (0.79), Lick_Mgb (0.77), Lick_Fe 5406 (0.72), D\(_{n}\)(4000) (0.74), g-r (0.76), r-i (0.74), u-z (0.76) |
| IC6 | Lick_Fe 5709 (-0.51)                                                 |
| IC11| EW (OIII 5007) (-0.64)                                               |
| IC14| \( \sigma\)-forbidden (0.38)                                        |

\[ \left( \begin{array}{c} 48 \\ 14 \end{array} \right) = 4.8 \times 10^{11} \] times for choosing the best set of ICs from the complete data set.
The remaining ICs are not dominated by any particular physical characteristic and the correlations with these ICs show negligibly small values. Thus even though we have got 14 ICs following IC1-IC14 (Subsections 3.2.1 & 3.3.2) among those, 6 ICs have negligibly small effects on the total variation and 8 of them are the most significant ones.

4.2. Properties of the galaxies in the groups

There are 11 effective groups (K1 - K8, K10 - K12), as a result of Cluster analysis. Among these K2, K4, K11 (viz. Table 2) consist of oldest galaxies with respect to average ages and $D_n(4000)$ values (though average age of K1 is higher than K2 but the average $D_n(4000)$ values are just the opposite and since it is an observed variables hence is more reliable one than the estimated value. Also K1 group of galaxies are similar to the groups of galaxies in the medium age range with respect to other physical properties e.g., color or metallicity etc.). K1, K3,K8, K10, K12 consist of galaxies of medium age, K6, K7 are the youngest groups of galaxies and K5 consists of unbarred galaxies. In each group we have classified 3 subgroups as strong barred (where the size of the bars are at least 30% of their host galaxy size), weak barred (where the size of the bars is smaller than 30% of the size of the host galaxy) and unbarred galaxies. It is clear from Table 2 that galaxies in groups K2, K4 and K11 have galaxies of median ages, and $D_n(4000)$ values which increase as, $K2 < K11 < K4$ and the median bar lengths of these groups also increase as $K2 < K11 < K4$. The metallicities are higher in these galaxies. On the contrary for galaxies in the
Figure 6. (b). Same as Figure 6a but only for unbarred spiral galaxies (green dots).

Figure 6. (c). Same as Figure 6a but only for barred spiral galaxies (red dots).
groups K1, K3, K8, K10, K12 the median ages and $D_n(4000)$ values vary more or less in the medium range ($\sim$ median age, 1.4 Gyr − 1.6 Gyr) with their bar lengths smaller compared to the oldest groups of galaxies. Finally K6, K7 are the youngest groups of unbarred galaxies with minimum sample size and K5 is the youngest group (Median age $\sim$ 1.0 Gyr) containing barred and unbarred galaxies.

4.3. Emission line diagnostics and CM diagrams

Emission line ratios have been recommended by various authors (Baldwin, Phillips, and Terlevich 1981; Veilleux and Osterbrock 1987; Kauffmann et al. 2003) for qualitative classification of galaxies. According to the above scheme, scatter diagram of two emission line ratios ($\log(NII/H\alpha)$ and $\log(OIII/H\beta)$) are classified by equations of curves separating the different classes of starburst galaxies, AGN and LINERs (Figure 6). It is clear from Figure 6 that K2, K4 and K11 contain all types of galaxies and the starburst galaxies have a wide range of ionization ($\log(OIII/H\beta) \sim -4$ to 0.5). K1, K3, K8, K10 and K12 groups are primarily dominated by starburst with few AGN and LINERs, wheras K6, K7, K5 are populated by starburst galaxies.

These observations are more or less consistent with the groups. As K2, K4 and K11 are the oldest groups of galaxies and they contain both unbarred and barred galaxies. Star formation still occurs in some unbarred galaxies accompanied by AGN and LINERs which are oldest in ages. In galaxies of medium ages in the groups K1, K3, K8, K10 and K12, they are dominated by star forming galaxies rather than AGN or LINERs where star formation is being affected by the presence of bars (Athanassoula 1983; Buta and Combes 1996). This is also reflected in Figure 7 where
Figure 7. (b). Color–magnitude ($u - z$ vs. $U$) diagrams of Unbarred galaxies and for each of the groups K1 - K12 except K9.

Figure 7. (c). Color–magnitude ($u - z$ vs. $U$) diagrams of Barred galaxies and for each of the groups K1 - K12 except K9.
for the oldest groups (K2, K4, K11) the contours peak to redder color, medium aged groups (K1, K3, K8, K10, K12) concentrate around bluer zone and in particular in Figure 8 where for the strong barred galaxies the contours comparatively peak at redder zone contrary to weak barred galaxies which peak at bluer zone in the color magnitude (CM) diagrams.

The color magnitude diagrams in \((g - r, R)\) and color-color diagrams in \((g - r, u - r)\) (Figure 9a and b) for all galaxies (strong barred, weak barred and unbarred) show similar features as in Figure 8. One interesting feature in color-color diagram shows that in K2 and K11 the weak barred galaxies occupy the same redder region as the strong barred galaxies. Since these galaxies fall in the oldest groups of galaxies, it indicates that these weak barred galaxies were previously strong barred galaxies but the bars are getting dissolved, supporting a recurrent phenomenon of bar formation. In the medium aged groups of galaxies the effect is not very pronounced due to non-availability of data on weak barred galaxies.

4.4. Star formation efficiency

Our aim is to assess the effect of bar on various properties of galaxies. Therefore we use the specific star formation rate variables (Star Formation Efficiency, hereafter SFE) denoted by log(SFR/M_*) (viz. Table 1) as a good measure for star formation activity in galaxies (Brinchmann et al. 2004). Also D_n(4000) (Balogh et al. 1999; Kauffmann et al. 2003) has been used for the second estimate of age of stellar populations. Figure 10 shows the SFE and \(D_n(4000)\) vs log(M_*) for various groups K1 - K12. For the oldest groups i.e., in K2, K4, K11, data for K4 are not available.
For K2, low mass weak barred galaxies have lower SFE compared to strong barred and unbarred galaxies \( (M_* \leq 10^{10.5} M_{\odot}) \). This might indicate that these galaxies were initially strong barred and their bars are getting dissolved and their SFE got attenuated in the presence of strong bar initially. Thus they have low SFE (Vera, Alonso, and Coldwell 2016). High mass weak barred galaxies \( (M_* > 10^{10.5} M_{\odot}) \) have their bars growing from unbarred galaxies, which is quenching their star formation activity and they have again low SFE. Generally SFE is always lower in strong barred galaxies than unbarred galaxies. Various authors (Ho, Filippenko, and Sargent 1997; Sheth et al. 2005; Ellison et al. 2011; Lee et al. 2012; Masters et al. 2012) have suggested that the presence of bars funnel the gas into the central region of the galaxy. Subsequently the material turned into molecular gas which when becomes gravitationally unstable undergo fragmentation and trigger star formation activity. Therefore presence of bar could accelerate gas consumption ceasing formation of new stars in the outer region of the disks as they become redder. Thus in the star burst phase large amount of gas is transported toward the galactic central region for triggering star formation activity and in the post star burst phase the gas is consumed by circum nuclear star burst which shows low star formation rate (SFR) (Jogee, Scoville, and Kenney 2005; Sheth et al. 2005).

For K2, low mass weak barred galaxies have lower SFE compared to strong barred and unbarred galaxies \( (M_* \leq 10^{10.5} M_{\odot}) \). This might indicate that these galaxies were initially strong barred and their bars are getting dissolved and their SFE got attenuated in the presence of strong bar initially. Thus they have low SFE (Vera, Alonso, and Coldwell 2016). High mass weak barred galaxies \( (M_* > 10^{10.5} M_{\odot}) \) have their bars growing from unbarred galaxies, which is quenching their star formation activity and they have again low SFE. Generally SFE is always lower in strong barred galaxies than unbarred galaxies. Various authors (Ho, Filippenko, and Sargent 1997; Sheth et al. 2005; Ellison et al. 2011; Lee et al. 2012; Masters et al. 2012) have suggested that the presence of bars funnel the gas into the central region of the galaxy. Subsequently the material turned into molecular gas which when becomes gravitationally unstable undergo fragmentation and trigger star formation activity. Therefore presence of bar could accelerate gas consumption ceasing formation of new stars in the outer region of the disks as they become redder. Thus in the star burst phase large amount of gas is transported toward the galactic central region for triggering star formation activity and in the post star burst phase the gas is consumed by circum nuclear star burst which shows low star formation rate (SFR) (Jogee, Scoville, and Kenney 2005; Sheth et al. 2005).

In K4, SFE values are not available for barred galaxies. In K11, for high mass range \( (10^{10} M_{\odot} > M > 10^9 M_{\odot}) \) strong barred and weak barred galaxies have similar SFE, which may be as low as \( \sim 10^{-11} \text{ yr}^{-1} \). Few unbarred galaxies have very low SFE \( (\sim 10^{-11} \text{ yr}^{-1}) \). Since they are of oldest ages their SFEs are low. Few weak barred galaxies in the high mass range have low SFE. This is indicative of the fact that these galaxies had strong bar initially but is getting dissolved. So, the SFE is low in these weak-barred galaxies. Some of this fact is consistent with the recurrent bar
formation scenario in galaxies (Bournaud and Combes 2002; Berentzen et al. 2004; Kormendy and Kennicutt Jr 2004; Gadotti and de Souza 2006; Katz et al. 2018; de Lorenzo-Cáceres et al. 2019; Hilmi et al. 2020).

In the medium aged groups of galaxies K1, K3, K12, they have similar phenomenon of recurrent bar formation. e.g., in K1, K3 and K12 some features have been observed as in case of oldest groups.

4.5. Mass-metallicity relation

In Figure 11 the mass-metallicity relation is shown for all the groups (K1 - K12 except K9). In the oldest groups of galaxies (K2, K4, K11), barred galaxy metallicities are always larger than unbarred ones. Metallicities increase with galaxy mass but the increase is almost constant after $z > 0.02$. For the same metallicity barred galaxy masses are higher than unbarred ones. In these groups the SFE decreases with increasing mass. This might be due to the fact that (Ellison et al. 2008, 2011) metal enhancement without an accompanying increase in star formation activity may be due to a short lived phase of bar-triggered star formation in the past. Also the fall in the metallicity is rapid for strong barred galaxies rather than weak or unbarred galaxies. Weak barred galaxy curve crossed the strong bar curve in the high mass zone as well as in low mass range. This might be due to the fact that in the low mass range they can grow from strong barred ones as their bars are getting dissolved and from unbarred ones, in the high mass zone, when SFEs are low (viz. Figure 10). For the intermediate aged groups (K1, K3, K8, K10, K12) in most cases the data are not available from sdss but the more-or-less trend is similar as that of oldest groups.
Figure 9. (a). Color distributions $(g - r)$ are plotted against $R$ values for the groups K1 - K12 except K9 where the yellow dots indicate unbarred galaxies, red dots indicate strong barred galaxies and the blue dots indicate weak barred galaxies.

Figure 9. (b). Color distributions $(g - r)$ are plotted against $(u - r)$ values for the groups K1 - K12 except K9 where the yellow dots indicate unbarred galaxies, red dots indicate strong barred galaxies and the blue dots indicate weak barred galaxies.
Moreover the crossing of weak bar curve with strong and unbarred ones indicates recurrent phenomenon of bar formation.

Figure 10. (a). SFE (yr⁻¹) values are plotted against log($M_\star$) ($M_\odot$) for the groups K1 - K12 except K9 where the yellow line is for unbarred galaxies, red line is for strong barred galaxies and the blue line is for weak barred galaxies.

Figure 10. (b). $D_n4000$ values are plotted against log($M_\star$) ($M_\odot$) for the groups K1 - K12 except K9 where the yellow line is for unbarred galaxies, red line is for strong barred galaxies and the blue line is for weak barred galaxies.
5. Conclusion

The present work deals with a large data set of unbarred, strong barred and weak barred galaxies taken from sdss DR15 and cross-matched with zooniverse, for collecting bar properties of the barred galaxies. We have considered several significant observable variables (e.g., Lick indices, Metallicity, SFR, Color Magnitudes etc.) (Table 1) for performing the statistical analyses and the estimated variables (e.g., age, SFE etc) along with observable ones are used for physical interpretation of the homogeneous groups. We have studied the influence of bars on the various properties of spiral galaxies. The following conclusions have been drawn from the above study:

1. The entire data set is classified into 12 homogeneous groups among which group 9 is an outlier. Instead of considering all the variables, we have classified the data set with respect to 14, ICs which are linear combinations of various variables with different weights. This is suitable for a non-Gaussian data set in a multivariate set up. We have found 12 groups by two independent methods, K-means cluster analysis (CA) and Gaussian Mixture Model Based Clustering (GMMBC) with respect to 14, ICs which establishes the robustness of the classification.

2. We have 14 Independent Components among which 8 components are found to be significant and they represent various influential physical galaxy properties like metallicity, ionization, color, absorption features and velocity dispersion etc. Remaining 6 components do not carry much variation in the galaxy properties. With respect to these 14 components the data set has been classified into 12 homogeneous groups by K-means clustering and the robustness has been established by another widely used method GMMBC.

3. Among these 12 groups, four groups (K2, K4, K11) fall in the oldest age (~ 2.6 Gyr - 6.75 Gyr) category and the groups (K1, K3, K8, K12) fall in the medium range (~ 1.68 Gyr - 1.8
Gyr). One group is an outlier (K9) and the remaining two groups (K6 - K7) are the youngest smallest groups of unbarred galaxies.

4. Oldest groups have longest bar lengths, highest metallicities and fall in the redder zone of color-magnitude diagram.

5. Galaxies of medium age range have shorter bar lengths and the groups are predominated by star burst galaxies, showing a kind of bluer zone.

6. In particular, weak barred galaxies show indication of recurrent bar formation scenario when the color-magnitude, color-color diagrams are studied thoroughly. This is consistent with the theoretical works suggested by various authors (Kormendy and Kennicutt Jr (2004) and Gadotti and de Souza (2006).

7. It has been found that presence of bars may affect the SFE, metallicity, color magnitude and nature of galaxies. When the barred galaxies are oldest, they are redder, their bar lengths are longer and SFE are lower. Few weak barred galaxies, which are precursors of strong barred galaxies, have lower SFE and few weak barred galaxies which are precursors of unbarred galaxies of lower masses have higher SFE. This is the very new feature reflected in the present study and concludes that bar formation is not always one way phenomenon but may get dissolved in course of time in oldest and medium age ranged galaxies.

Acknowledgement

The authors are grateful to the sdss Data Release 15 (sdss DR15) for giving the opportunity to retrieve the data and to work with it. One of the authors Prasenjit Banerjee acknowledges to a research fellowship provided by The University Grants Commission, with UGC Id. NOV2017-422665 & UGC Reference No. 1141/(CSIR-UGC Net June 2017).

References

Agarwal, S., J. Lim, L. Zelnik-Manor, P. Perona, D. Kriegman, and S. Belongie. 2005. Beyond pairwise clustering. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05). Vol. 2, 838–45, IEEE.

Aguerri, J. A. L., and A. C. González-García. 2009. On the origin of dwarf elliptical galaxies: The fundamental plane. Astronomy & Astrophysics 494 (3):891–904. doi:10.1051/0004-6361:200810339.

Albazzaz, H., and X. Z. Wang. 2004. Statistical process control charts for batch operations based on independent component analysis. Industrial & Engineering Chemistry Research 43 (21):6731–41. doi:10.1021/ie049582+.

Almeida, J. S., J. A. L. Aguerri, C. Munoz-Tunón, and A. De Vicente. 2010. Automatic unsupervised classification of all Sloan digital sky survey data release 7 galaxy spectra. The Astrophysical Journal 714 (1):487–504. doi:10.1088/0004-637X/714/1/487.

Alonso, M., G. Coldwell, and D. G. Lambas. 2013. Effect of bars in AGN host galaxies and black hole activity. Astronomy & Astrophysics 549:A141. doi:10.1051/0004-6361/201220117.

Alonso, S., G. Coldwell, and D. G. Lambas. 2014. AGN spiral galaxies in groups: Effects of bars. Astronomy & Astrophysics 572:A86. doi:10.1051/0004-6361/201424523.

Athanassoula, E. 1983. Formation of rings and lenses. Symposium - International Astronomical Union 100:243–52. doi:10.1017/S007418090003299X.

Athanassoula, E. 2003. What determines the strength and the slowdown rate of bars? Monthly Notices of the Royal Astronomical Society 341 (4):1179–98. doi:10.1046/j.1365-8711.2003.06473.x.

Athanassoula, E. 2012. Manifold-driven spirals in N-body barred galaxy simulations. Monthly Notices of the Royal Astronomical Society: Letters 426 (1):L46–L50. doi:10.1111/j.1745-3933.2012.01320.x.

Athanassoula, E., J. Lambert, and W. Dehnen. 2005. Can bars be destroyed by a central mass concentration?—I. Simulations. Monthly Notices of the Royal Astronomical Society 362 (2):496–508. doi:10.1111/j.1365-2966.2005.09445.x.

Babu, G. J., T. Chattopadhyay, A. K. Chattopadhyay, and S. Mondal. 2009. Horizontal branch morphology of globular clusters: A multivariate statistical analysis. The Astrophysical Journal 700 (2):1768–78. doi:10.1088/0004-637X/700/2/1768.

8https://skyserver.sdss.org/dr15/en/tools/search/sql.aspx
Martinez-Valpuesta, I., L. Shlosman, and C. Heller. 2006. Evolution of stellar bars in live axisymmetric halos: Recurrent buckling and secular growth. The Astrophysical Journal 637 (1):214–26. doi:10.1086/498338.

Martin, P., and J.-R. Roy. 1994. The influence of bars on the chemical composition of spiral galaxies. The Astrophysical Journal 424:599–614. doi:10.1086/173917.

Martins-Filho, W., C. Griffith, K. Pearson, I. Waldmann, A. Alvarez-Candal, and R. T. Zellem. 2018. Independent component analysis applied to ground-based observations. American Astronomical Society Meeting Abstracts 231:148–26.

Masters, K. L., R. Nichol, S. Bamford, M. Mosleh, C. J.Lintott, D. Andreescu, E. M. Edmondson, W. C. Keel, P. Murray, M. J. Raddick, et al. 2010. Galaxy Zoo: Dust in spiral galaxies. Monthly Notices of the Royal Astronomical Society 404 (2):792–810. doi:10.1111/j.1365-2966.2010.16335.x.

Masters, K. L., R. C. Nichol, M. P. Haynes, W. C. Keel, C. Lintott, B. Simmons, R. Skibba, S. Bamford, R. Giovanelli, and K. Schawinski. 2012. Galaxy Zoo and ALFALFA: Atomic gas and the regulation of star formation in barred disc galaxies. Monthly Notices of the Royal Astronomical Society 424 (3):2180–92. doi:10.1111/j.1365-2966.2012.21377.x.

Modak, S., T. Chattopadhyay, and A. K. Chattopadhyay. 2017. Two phase formation of massive elliptical galaxies: Study through cross-correlation including spatial effect. Astrophysics and Space Science 362 (11):206. doi:10.1007/s10509-017-3171-3.

Modak, S., T. Chattopadhyay, and A. K. Chattopadhyay. 2020. Unsupervised classification of eclipsing binary light curves through k-medoids clustering. Journal of Applied Statistics 47 (2):376–92. doi:10.1080/02664763.2019.1635574.

Mondal, S., A. K. Chattopadhyay, and T. Chattopadhyay. 2008. Globular clusters in the Milky Way and dwarf galaxies: A distribution-free statistical comparison. The Astrophysical Journal 683 (1):172–7. doi:10.1086/589851.

Mu, B. 2007. Unsupervised spectral classification of astronomical x-ray sources based on independent component analysis. Thesis. Rochester Institute of Technology.

Murtagh, F., and A. Heck. 2012. Multivariate data analysis. Vol. 131. Springer Science & Business Media.

Newnham, L., K. M. Hess, K. L. Masters, S. Kruk, S. J. Penny, T. Lingard, and R. Smethurst. 2020. The H i morph and stellar properties of strongly barred galaxies: Support for bar quenching in massive spirals. Monthly Notices of the Royal Astronomical Society 492 (4):4697–715. doi:10.1093/mnras/staa064.

Norman, C. A., J. Sellwood, and H. Hasan. 1996. Bar dissolution and bulge formation: An example of secular dynamical evolution in galaxies. The Astrophysical Journal 462:114. doi:10.1086/177133.

Oh, S., K. Oh, and K. Y. Sukyoung. 2012. Bar formation on central star formation and active galactic nucleus activity. The Astrophysical Journal Supplement Series 198 (1):4. doi:10.1088/0067-0049/198/1/4.

Peth, M. A., J. M. Lotz, P. E. Freeman, C. McPartland, S. A. Mortazavi, G. F. Snyder, G. Barro, N. A. Grogin, Y. Guo, S. Hemmati, et al. 2016. Beyond spheroids and discs: Classifications of CANDELS galaxy structure at 1.4 < z < 2 via principal component analysis. Monthly Notices of the Royal Astronomical Society 458 (1):963–87. doi:10.1093/mnras/stw252.

Pettitt, A. R., and J. W. Wadsley. 2018. Bars and spirals in tidal interactions with an ensemble of galaxy mass models. Monthly Notices of the Royal Astronomical Society 474 (4):5645–71. doi:10.1093/mnras/stx3129.

Pfister, N., S. Weichwald, P. Bühlmann, and B. Schölkopf. 2019. Robustifying independent component analysis by adjusting for group-wise stationary noise. Journal of Machine Learning Research 20 (147):1–50.

Pike, S., K. Ebisawa, S. Ikeda, M. Morii, M. Mizumoto, and E. Kusunoki. 2017. Application of data science techniques to disentangle X-ray spectral variation of super-massive black holes. arXiv preprint arXiv:1701.05386.

Pires, S., J. Juin, D. Yvon, Y. Moudden, S. Anthoine, and E. Pierpaoli. 2006. Sunyaev-Zel’dovich cluster reconstruction in multiband bolometer camera surveys. Astronomy & Astrophysics 455 (2):741–55. doi:10.1051/0004-6361:20053820.

Pompea, S. M., and G. H. Rieke. 1990. A test of the association of infrared activity with bars. The Astrophysical Journal 356:416–29. doi:10.1086/168849.

Rao, C. R., C. RadhakrishnaRao, J. Kleffe, and J. Kleffe. 1988. Estimation of variance components and applications. Vol. 3. North Holland.

Richardson, C. T., J. T. Allen, J. A. Baldwin, P. C. Hewett, G. J. Ferland, A. Crider, and H. Meskhidze. 2016. Interpreting the ionization sequence in star-forming galaxy emission-line spectra. Monthly Notices of the Royal Astronomical Society 458 (1):988–1012. doi:10.1093/mnras/stw100.

Roberts, W. W., Jr, J. M. Huntley, and G. D. van Albada. 1979. Gas dynamics in barred spirals-Gaseous density waves and galactic shocks. The Astrophysical Journal 233:67–84. doi:10.1086/157367.

Rosas-Guevara, Y., S. Bonoli, M. Dotti, T. Zana, D. Nelson, A. Pillepich, L. C. Ho, D. Izquierdo-Villalba, L. Hernquist, and R. Pakmor. 2019. The buildup of strongly barred galaxies in the TNG100 simulation. Monthly Notices of the Royal Astronomical Society 491 (2):2547–64. doi:10.1093/mnras/stz3180.

Rousseeuw, P. J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20:53–65. doi:10.1016/0377-0427(87)90125-7.
