Machine Learning Estimation of Physical Properties of S0 Galaxies from their Optical Spectra

Author: Roger Almasqué Vila
*Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.**

Advisors: Josep Maria Solanes (UB), Jaime Perea (IAA-CSIC)

Abstract: In this work, we show that it is possible to infer precise information about some of the main physical properties of lenticular galaxies from the shape of their entire optical spectrum. We study this methodology as an alternative to the more conventional way of individually analyzing the most important emission and/or absorption lines in this frequency band. By using neural networks trained with high signal-to-noise spectra ranging from 400 nm to 800 nm, we have determined the accuracy of the predictions for the following interesting properties: the equivalent width (EW) of the emission lines H α, H β, [O III] and [N II]; the $D4000$ break, the specific star formation rate, sSFR, and the stellar mass to light ratio in the SDSS r-band, M_*/L_r. We provide a comparison of the performance of this method using as input, on the one hand, all the dimensionality available in the spectra and, on the other hand, only their first principal components (PC). We conclude that the latter procedure produces better results when predicting the selected variables. We have also inferred that 5 is the ideal number of PCs to compute the values of these variables and identified the most dominant ones to determine which and how many eigenspectra are required for a minimal optimal prediction. Finally, we have tested the performance of our methodology as a WHAN activity classifier, also obtaining encouraging results.

I. INTRODUCTION

Galaxies as distinct baryonic and dark matter units have many measurable properties that inform on both their current physical state and their evolutionary history. The properties that can be directly inferred include: spectral energy distributions, spectral line fluxes and equivalent widths, metallicities, light profiles, colors, sizes, etc. From these observable quantities, one can derive additional physical information, such as the stellar, gaseous and total masses, mass-to-light ratios, the kinematics, and the current and past star formation rates. Each of these features reveals potentially important clues for how galaxies were created and have evolved across cosmic time.

Efforts to infer these key physical properties focus on both photometric and spectroscopic observations in one or more broad wavelength bands. In the latter case, however, the use of the data is usually reduced to the determination of the fluxes contained in a series of relevant emission and/or absorption lines with which a large part of the information that these elements carry is wasted. Another of the handicaps that the use of spectral lines entails is the need to define a baseline to be able to adequately extract the contribution of the continuum. The fact that there are different strategies to estimate the continuum emission such as the fitting of high-degree polynomials to several wavelengths near each line that are assumed to be free of emission/absorption features or the use of spectral population synthesis codes that attempt to reproduce the observed spectra either from a spectral library that spans a relevant range in stellar age and metallicity or making assumptions on the star formation and chemical enrichment history of the galaxy, makes the results user dependent.

The aim of this work is to analyze the feasibility of a simple method to predict physically interesting properties of galaxies from their spectra that takes into account the wealth of information contained in all their single usable data elements (i.e., each one of the N flux pixels) and that does not require any subjective input from the user. Our methodology is a Machine Learning (ML) technique that consists on training a neural network (NN) with high-quality optical spectra and the corresponding values of the property that we want to infer, and then determine if it can be used as a pipeline to accurately predict the values of such variable for any galaxy. We are going to compare the performance of two different approaches for the input of the spectral information. On the one hand, we will feed the NNs with the totality of the optical spectra. On the other hand, we will replace the spectra by a small number of their principal components (PC). It is well known that the application of the principal component analysis (PCA) technique to a set of spectra linearly transforms the data into a new coordinate system where most of the variance is contained in the first $M << N$ principal directions (eigenvectors)[1]. This allows the replacement of each individual spectrum by the eigenvalues that result from its projection on a few principal directions, thus drastically reducing the dimensionality of the original dataset while preserving most of the information.

These two methodologies will be tested using the spectra of a subset of the database of nearby ($z < 0.1$) lenticular (or S0) galaxies defined in Tous et al. (2020)[2],

*Electronic address: ralmasvi67@alumnes.ub.edu

for which their eigenspectra are already known. Furthermore, the S0s have has properties intermediate between the large taxonomic families that constitute the two extremes of the Hubble sequence: they posses the general disk/bulge morphology of spiral galaxies, as well as baryonic contents usually dominated by old stars and with little warm and cold nebular components, representative of elliptical galaxies. Thus, we expect that the outcome of the present work can be transferable to these other morphologies.

The remaining sections of the paper are devoted to determine and compare the performance of the two realizations of our methodology. First, by applying them to the prediction of the values of a small set of seven different galaxy properties and then, to the more complex situation that represents a classifying prediction, equivalent to simultaneously predict two properties such as the EW(H α) and the [N II]/H α flux ratio that are required by the WHAN diagram (Cid Fernandes et al. 2010)[3] to classify galaxies according to their level of activity.

II. INDIVIDUAL PROPERTIES

Among the properties selected to investigate the quality of our ML-based predictions we have chosen the equivalent width (EW) of the H α, H β, [O III]λ5007Å and [N II]λ6584Å lines. The EW of a spectral emission/absorption line is a measure that synthesizes in a single number the strength of the line in relation to the underlying continuum level. In this sense, it provides a better measure of the relevance of the line than the total flux. The EW(H β) and especially EW(H α), are crucial for characterizing the star formation in galaxies. They are produced by the UV radiation field of massive newborn O-stars that ionises the gas surrounding them. The subsequent hydrogen recombination leads to emission lines in different wavelength ranges, mainly creating the Balmer series (H α, H β, ...). For their part, the forbidden [O III] and [N II] lines are are among the most prominent emission lines present in the spectra of both photoionized star-formimg nebulae and photoionized gas of the Narrow Line Region (NLR) surrounding Active Galactic Nuclei (AGN). Both lines are sensitive to the electron temperature.

More "complex" physical variables such as $D4000$, M_*/L_r, and $\log(sSFR)$ have been selected as well. The $D4000$ break, is a dimensionless parameter defined as the ratio of the flux densities in the range of 4050Å and 4250Å and the range of 3750Å and 3950Å, and is widely used to determine the stellar population age of galaxies. The ratio M_*/L_r is the relation between the galaxy's stellar mass and its total luminosity in the r (red) band of the Sloan Digital Sky Survey (SDSS) filter system. Finally, $\log(sSFR)$ characterizes how the instantaneous star formation is determined by the galaxy past formation history. It is well known that the SFR tightly correlates with stellar mass for star-forming galaxies, a relationship that is referred to as the Galaxy Main Sequence.

III. DATA AND ANALYSIS PIPELINE

The sample of single-fiber spectra we have used is a subset of the 68,000 spectral observations of S0 galaxies extracted by Tous et al (2020) from the twelve Data Release of the SDSS (SDSS-DR12; Alam et al. 2015 [4]) and that were used to determine the principal axes of the optical spectra of this population. We have also retrieved from this database the r-band absolute Petrosian magnitudes, while the observed values of the remaining variables have been gathered from two sources: the fluxes and equivalent widths of the lines are from the Portsmouth catalog [5], while for the stellar mass and sSFR we have used the values listed in the GALEX-SDSS-WISE Legacy Catalog 2 [6].

All spectra have gone through a process of imputation using the KNNimputer() with 5 nearest neighbours weighted by distance in order to fill any possible gap in them, followed by a cleaning procedure that have discarded any spectrum containing more than 10 per cent of pixels affected by sky lines or very large errors.

In this part of the work, the NN has been trained and tested with the spectra with high signal-to-noise ratio (S/N) and significant H α emission lines. We have done this by adopting a lower threshold for the amplitude over noise (AoN) of 1.5, following the indications of the Portsmouth survey, and a minimum $EW(\mathrm{H}\,\alpha)$ of 1.0. After applying these constraints our final sample consists of 27,134 galaxies.

The NN used in this work is based on Scikit-Learn MLPRegressor, that uses a stochastic gradient optimizer for the squared error and ReLU as activation function. When using the PCs as predictors, we build a network of 2 hidden layers with 20 neurons each for the line widths, and of 2 hidden layers with 80 neurons for $D4000$, M_*/L_r and $\log(sSFR)$. Extra layers or more neurons not only do not provide any improvement (see Table I) but rather increase the training time and the computing resources used, but also negatively affect the quality of the predictions. To compare the accuracy of the predictions, we use the score (R^2), defined as:

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2} \ , \tag{1}$$

where y_i are the observed value of the variables and f_i are the predictions made by the pipeline. Note that a score of 1 is a perfect prediction, while systematically predicting for a variable just the average of the observed values, $\bar{y}$, would give a score of 0. For worse predictions, the score can be arbitrarily negative. We show in Table I the CPU times and scores achieved by predicting the $EW(\mathrm{H}\,\alpha)$ from 5 PCs and different configurations of the NN.

All the computations in this work have been performed remotely on a Xeon 6138 machine with 155 GB of RAM.

Hidden layers	Neurons	Time (s)	Score
2	20	1.39 ± 0.02	0.985 ± 0.001
2	10	2.10 ± 0.04	0.983 ± 0.001
2	30	1.44 ± 0.08	0.983 ± 0.001
3	30	1.39 ± 0.02	0.984 ± 0.001
4	40	3.48 ± 0.09	0.984 ± 0.001

TABLE I: CPU times and scores for the prediction of $EW(\mathrm{H}\,\alpha)$ with 5 PCs and different NN configurations.

CPU times may change significantly with other hardware, but the score will remain unchanged. In the same way, it has been found that 3 hidden layers with 50 neurons each give the best performance when the whole spectrum is used as input. Similarly, a more complex structure of 4 hidden layers with 120 neurons provides the best performance for the properties $D4000$, M_*/L_r and $\log(sSFR)$.

The type of NN we are using does not assign an uncertainty to the predicted values. So, in order to get a rough estimate of the error of the predictions, we have assumed that the studied variables follow normal distributions centered in their observed values and with a scale equal to their associated errors and computed the standard deviation of the predictions that result from randomly selecting 200 values from them.

IV. RESULTS FOR SINGLE PROPERTIES

A. PCs vs global spectrum predictions

The main goal of this paper is to determine and compare the accuracies of the predictions for physical variables of galaxies obtained by means of ML pipelines that use either the PCs of the optical spectra or the full spectra as input. There are evident benefits using PCs: less memory usage, computational cost and invested time. None of them will be relevant if it is not accompanied by sufficiently precise results. For the benchmarking, we have generated ML pipelines for both types of input, limiting the PC-based ones to the first 5 PCs, because there is no significant gain in the accuracy when increasing their number (see Secs. IV.b and IV.c).

As Table II shows, the PC-based pipelines perform better that their full spectrum counterparts for all the variables studied. In fact, for $D4000$, M_*/L_r and $\log(sSFR)$ we obtain a minimally acceptable score only when using PCs. The error quoted for the scores has been inferred assuming a normal distribution with no correlation, using ten random training sets as sample. Fig. 1 allows to display the improvement for D4000 that results from using PC-based pipelines. A similar behavior is observed for the rest of the variables.

	Principal components	Full spectrum
$EW(\mathrm{H}\,\alpha)$	0.985 ± 0.001	0.958 ± 0.002
$EW(\mathrm{H}\,\beta)$	0.948 ± 0.001	0.918 ± 0.002
$EW([\mathrm{O\,III}])$	0.982 ± 0.001	0.956 ± 0.002
$EW([\mathrm{N\,II}])$	0.948 ± 0.003	0.924 ± 0.005
$D4000$	0.934 ± 0.002	0.53 ± 0.07
M_*/L_r	0.674 ± 0.002	0.49 ± 0.04
$\log(sSFR)$	0.750 ± 0.005	-4.2 ± 0.6

TABLE II: Scores of the PC- and full-spectrum-based NNs for each of the selected variables.

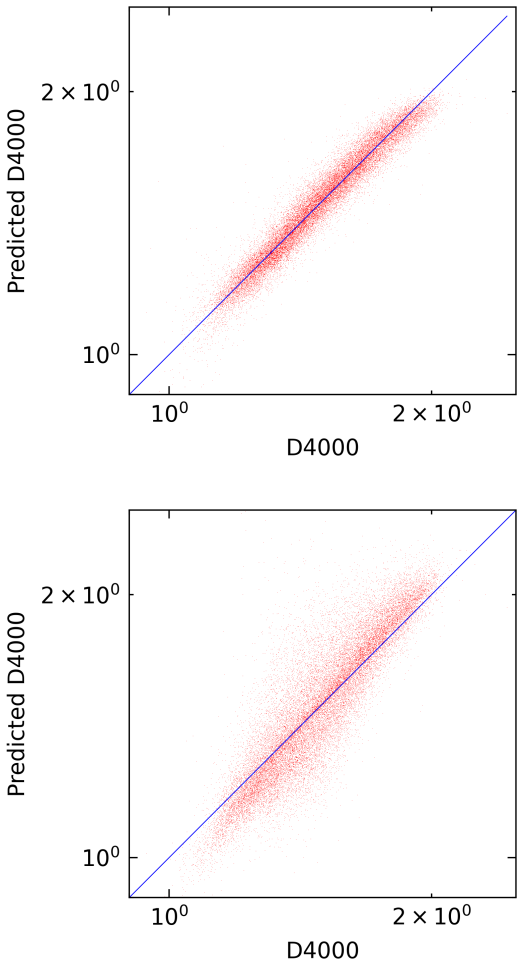

FIG. 1: Predicted vs measured values of D4000. Top: using the first 5 PCs of the optical spectrum as input. Bottom: using the full spectrum as input. The blue line represents the ideal result $y = x$.

Num. of PCs	$EW(\mathrm{H}\,\alpha)$	$EW(\mathrm{H}\,\beta)$	$EW([\mathrm{O\,III}])$	$EW([\mathrm{N\,II}])$	$D4000$	M_*/L_r	$\log(sSFR)$
1	0.857	0.830	0.363	0.580	0.816	0.554	0.639
2	0.967	0.943	0.430	0.800	0.894	0.677	0.716
3	0.983	0.948	0.959	0.837	0.910	0.676	**0.751**
4	0.984	0.948	**0.981**	0.918	0.927	0.671	0.747
5	0.984	0.949	0.981	0.946	**0.932**	0.671	0.744
6	0.989	0.957	0.981	0.953	0.916	0.689	0.747
7	0.989	0.958	0.981	0.963	0.921	**0.691**	0.738
8	**0.990**	**0.962**	0.980	0.965	0.872	0.688	0.750
9	0.989	0.961	0.954	**0.968**	0.867	0.691	0.733
10	0.989	0.960	0.976	0.965	0.866	0.677	0.720

TABLE III: Scores of the NN predictions using different number of PCs as input. The maximum scores for each property are highlighted in boldface. The estimated errors of the listed values are all in the range 0.001-0.005.

$EW(\mathrm{H}\,\alpha)$		$EW(\mathrm{H}\,\beta)$		$EW([\mathrm{O\,III}])$		$EW([\mathrm{N\,II}])$		$D4000$		M_*/L_r		$\log(sSFR)$	
1	24.5%	2	27.1%	3	29.0%	5	22.5%	5	23.5%	2	23.7%	3	20.8%
2	23.7%	1	21.7%	2	22.2%	1	22.1%	1	22.6%	5	20.9%	1	20.3%
3	19.2%	3	18.7%	1	20.6%	4	19.6%	3	19.5%	3	20.1%	5	20.1%
5	16.5%	4	18.6%	4	15.6%	3	19.2%	4	18.4%	4	18.3%	4	19.6%
4	16.1%	5	14.0%	5	12.6%	2	16.6%	2	16.0%	1	17.0%	2	19.2%

TABLE IV: Percentual contributions to the predictions (right col.) of the first 5 PCs (left col.) in decreasing order.

B. Finding the ideal number of PCs

By increasing the number of PCs in the NNs, we are getting closer to the performance of the full spectrum. Therefore, there must be for each of the adopted variables an ideal number of PCs for which the accuracy of the predictions reaches its maximum. To determine the number of PCs that maximize the accuracy we have built pipelines that use up to ten eigenvectors as input.

As shown in Table III, only 5 PCs are needed for the scores of all the variables investigated to reach values close to their respective maximum precision. Increasing the number of PCs beyond this value does not significantly improve the predictions, with even a deterioration in precision being observed in cases where predictions based on the full spectrum perform poorly, i.e., for $\log(sSFR)$, M_*/L_r, and, especially, $D4000$. Note also that the predictions for these three variables require only one PC to outperform the full spectrum predictions, while $EW(\mathrm{H}\,\alpha)$ and $EW(\mathrm{H}\,\beta)$ require two, $EW([\mathrm{O\,III}])$ three, and $EW([\mathrm{N\,II}])$ five.

C. Dominant PCs

An alternative way to infer the optimal number of PCs is to determine which eigenvalues contain most of the information of the studied variables. To achieve this goal, we have built a boosting framework by means of the tool LightGBM that uses gradient booster decision trees to calculate the weights of the different components that contribute to a given result. In Table IV, we show for each variable the relative percentage weights of the contributions of the first 5 eigenspectra. It can be seen that the predictions for the EWs of the two Balmer lines rely on either first two PCs or in both (Ha), while for EW([OIII]) and log(sSFR) most of the information is encapsulated into the third eigenvalue, although in the latter case there is no clearly dominant component. In contrast, the highest weight of EW([N II]) and $D4000$ is carried by the fifth component, although in both cases there is an important contribution of the first PC. These results confirm that predictive NNs based on the first 5 principal components of the optical spectrum are the most suitable regardless of the investigated property (at least for the set of selected variables).

V. RESULTS FOR THE WHAN CLASSIFIER

As an example of a more complex application of our methodology, we have built an activity classifier for galaxies based on the WHAN diagram. This scheme, that depicts the $EW(\mathrm{H}\,\alpha)$ versus the logarithm of the flux ratio $[\mathrm{N\,II}]/\mathrm{H}\,\beta$, is capable of dealing with galaxies with weak o absent emission lines, therefore increasing the census of objects qualifying for classification, while offering at the same time a similar o even better diagnostic power than classical activity diagrams.

In the WHAN diagram, active galaxies, i.e., those with strong ionization emission lines, are subdivided into star-forming galaxies (SF) and galaxies with an active galactic nucleus (AGN), with the latter divided into Seyfert or strong AGN (sAGN) and LINER or weak AGN (wAGN). Galaxies with weaker emission lines are subdivided into two groups: emission-line retired galaxies (RG), that show nebular emission lines arising from photoionisation

by post-AGB stars, and line-less retired or truly passive galaxies (PG) with undetected H α emission. The demarcations of these five classes are [7]:

SF: $\log(\frac{[\mathrm{N\,II}]}{\mathrm{H}\,\alpha}) < -0.4 \;\;\&\;\; EW(\mathrm{H}\,\alpha) > 0.5\text{Å}$

sAGN: $\log(\frac{[\mathrm{N\,II}]}{\mathrm{H}\,\alpha}) < -0.4 \;\;\&\;\; EW(\mathrm{H}\,\alpha) > 6\text{Å}$

wAGN: $\log(\frac{[\mathrm{N\,II}]}{\mathrm{H}\,\alpha}) < -0.4 \;\;\&\;\; 3\text{Å} < EW(\mathrm{H}\,\alpha) < 6\text{Å}$

RG: $\log(\frac{[\mathrm{N\,II}]}{\mathrm{H}\,\alpha}) < -0.4 \;\;\&\;\; 0.5\text{Å} < EW(\mathrm{H}\,\alpha) < 3\text{Å}$

PG: $EW(\mathrm{H}\,\alpha) < 0.5\text{Å} \;\;\&\;\; EW([\mathrm{N\,II}]) < 0.5\text{Å}$

Note that a good property of activity classification diagrams like WHAN is that their are essentially insensitive to the effects of dust extinction. This is achieved by using line ratios that involve emission lines with similar wavelengths ([N II] = 6548Å, H α = 6563Å) and/or EWs that compare the flux of a narrow emission line with the underlying continuum. Nevertheless, the computation of the line fluxes and equivalent widths requires adopting a certain theoretical stellar evolution model to subtract the continuum. This leads to discrepancies in the results from different authors (larger when flux ratios are involved) that may affect the classification of some galaxies.

For this reason, it is interesting to test the feasibility of building a user-independent estimator of the WHAN activity classes from a NN that uses the optical spectra of the galaxies as input. To carry out this endeavour we have defined a new sample of nearby S0 galaxies by applying the $AoN > 1.5$ condition to the Tous et al. (2020) dataset, but removing the lower threshold filter on $EW(\mathrm{H}\,\alpha)$ to allow for the incorporation of galaxies with weak or no activity. This has resulted in a larger sample with a total of 45,049 objects, of which once again an eighth part has been used as a training set.

We have found that the best performance is achieved for 2 hidden layers with 20 neurons each, producing an acceptable score of 0.845 with 5 PCs as input and reaching a maximum score of 0.902 with 7 PCs. Less promising results are obtained with the complete optical spectra, reaching a maximum score of 0.727 with 3 hidden layers and 80 neurons each.

VI. CONCLUSIONS

In this work, we have presented a new and successful machine learning-based methodology for the prediction of the most important physical properties of galaxies, only using both their full rest-frame single-fiber optical spectrum and its first principal spectral components (PCs) as input in a neural network (NN). The performance of this methodology has been tested on sample of spectra from nearby lenticular galaxies extracted from the SDSS. For these galaxies, we have predicted the equivalent width of the H α, H β, [O III], and [N II] lines, the $D4000$ break, the specific star formation rate, and the stellar mass to red light ratio, as well as their WHAN activity classes. In all these cases, the use of the PCs has always led to more accurate results. Dealing with PCs also has the advantage of being much less CPU time-consuming than using full spectra, especially during the NN training stage.

We have also investigated the quality of the predictions based on the number of PCs, as well as identified which PCs most decisively affect the selected parameters, concluding that five would be an ideal number (at least for the variables studied). We do not expect that the results of this work depend substantially on the source of the spectra, as long as their spectral resolution is similar to that of the Sloan spectrograph, nor on the morphology of the galaxies that are analyzed.

Possible future developments of this thesis involve the application of performance tests to other physical properties, such as the metallicity, $[Fe/H]$, and the abundance of alpha elements, $[\alpha/Fe]$, upgrading the predictor pipeline by introducing a more realistic determination of the predictions error through the Markov chain Monte Carlo method, and extending this methodology to other Hubble types.

Acknowledgments

I would like to thank my supervisors Josep Maria Solanes and, specially, Jaime Perea for giving me the opportunity to work with them in a real-world project, and also providing me the knowledge and assistance at every stage of the project.

[1] Sodré L., Cuevas H., *Global regularities in integrated galaxy spectra* MNRAS 287, 137 (1997)

[2] Tous, J.L., et al, *The local universe in the era of large surveys. I. Spectral classification of S0 galaxies.* MNRAS 495, 4135 (2020)

[3] Fernandes, R.C. et al, *Alternative diagnostic diagrams and the 'forgotten' population of weak line galaxies in the SDSS* MNRAS 403, 1036 (2010)

[4] S. Alam et al, *The eleventh and twelfth data releases of the SLOAN Digital Sky Survey: Final data from SDSS-III* ApJS 219, 12 (2015)

[5] D. Thomas et al, *Stellar velocity dispersions and emission line properties of SDSS-III/BOSS galaxies* MNRAS 431, 1383 (2013)

[6] S. Salim et al, *Dust Attenuation Curves in the Local Universe: Demographics and New Laws for Star-forming Galaxies and High-redshift Analogs* ApJ 859, 1 (2018)

[7] Fernandes, R.C., et al, *A comprehensive classification of galaxies in the Sloan Digital Sky Survey: how to tell true from fake AGN?* MNRAS 413, 187 (2011)