Statistical Challenges in 21st Century Cosmology
Proceedings IAU Symposium No. 306, 2014
A. F. Heavens, J.-L. Starck & A. Krone-Martins, eds.
© International Astronomical Union 2015
doi:10.1017/S1743921314010990

Quantifying correlations between galaxy emission lines and stellar continua using a PCA-based technique

Róbert Beck, László Dobos and István Csabai

Department of Physics of Complex Systems, Eötvös Loránd University,
Pf. 32, H-1518, Budapest, Hungary
E-mail: robert.beck23@gmail.com (RB), dobos@complex.elte.hu (LD),
csabai@complex.elte.hu (ICs)

Abstract. We analyse the correlations between continuum properties and emission line equivalent widths of star-forming and narrow-line active galaxies from SDSS. We show that emission line strengths can be predicted reasonably well from PCA coefficients of the stellar continuum using local multiple linear regression. Since upcoming sky surveys will make broadband observations only, theoretical modelling of spectra will be essential to estimate physical properties of galaxies. Combined with stellar population synthesis models, our technique will help generate more accurate model spectra and mock catalogues of galaxies to be used to fit data from new surveys. We also show that, by combining PCA coefficients from the pure continuum and the emission lines, a plausible distinction can be made between weak AGNs and quiescent star-forming galaxies. Our method uses a support vector machine, and allows a more refined separation of active and star-forming galaxies than the empirical curve of Kauffmann *et al.* (2003).

Keywords. Galaxies: emission lines, starburst, active, classification, Methods: data analysis

1. Data sample and preliminary processing

We selected a sample of 13834 galaxies from the SDSS Data Release 7, with both photometric and spectroscopic measurements. The galaxies had to contain 11 given emission lines, and had to have a signal to noise ratio larger than 5. The sample size was limited by allowing only a given range of right ascension.

We performed dereddening on the galaxy spectra, transformed them into the rest frame, and normalised them by dividing by the average of median flux values in given wavelength ranges. Then the spectra were subdivided into stellar continuum and emission line components. The continuum component was given as the linear combination of 10 model spectra, while the emission lines were fitted, and characterized by their equivalent widths.

The continuum components of the spectra were processed further, we performed principal component analysis (PCA) on them via singular value decomposition. We kept the first 5 principal components that explained the largest amount of deviation in the data. Hereafter the continuum properties of the spectra will be characterized by their five singular vector coefficients.

2. Local multiple linear regression

We tried to quantify the connection between the stellar continua and the emission lines. Thus, we performed local multiple linear regression in the PCA space of the continua,

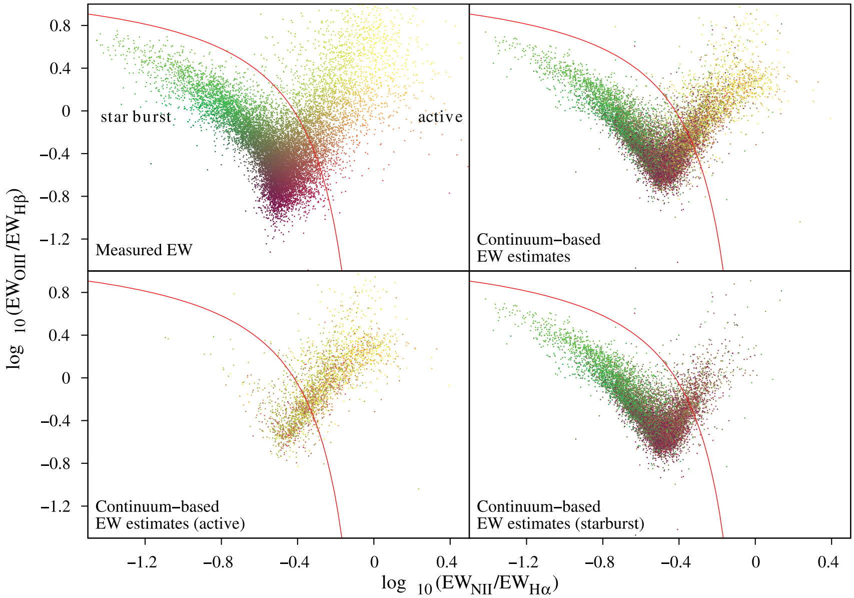

Figure 1. The BPT diagrams of our sample. The colouring corresponds to the galaxies' location on the top left image. The empirical separation line of Kauffmann *et al.* (2003) is shown with a solid line, the classification in the bottom images is based on that. The fitted equivalent widths follow the same V-shaped relationship, but with less deviation around it. It is visible that a significant number of starburst galaxies are estimated to be active based on their stellar continua.

fitting the logarithm of emission line equivalent widths. The 5 singular vector coefficients (that characterize the continua) were the independent variables, and the logarithm of the EWs were the dependent variables in the local regression. In other words, we expressed the line EWs locally as a linear function of continuum PCA coefficients. Locality was determined with a kD-tree, using Euclidean distance, and we took into account only the 30 nearest neighbours. We found that the emission line equivalent widths could be estimated reasonably well from the stellar continua using the method described above.

The lines' measured $\ln(EW)$ values were in the $[-2, 6]$ range, and the RMS values of the estimation (the square root of the mean squared error) were in the $[0.37, 0.78]$ interval, with Pearson's correlation coefficients in the $[0.60, 0.90]$ range. For 13834 data points, this yielded a p-value below floating point accuracy even in the worst case.

3. BPT diagrams

The Baldwin–Phillips–Terlevich diagram (Baldwin *et al.* 1981) is used to quickly characterize a population of galaxies based on emission line properties. The empirical separation line of Kauffmann *et al.* (2003) distuinguishes between active galaxies and star-forming galaxies:

$$\log_{10}\frac{[OIII]}{H\beta} > \frac{0.61}{\log_{10}\left([NII]\,/H\alpha\right) - 0.05} + 1.3$$

If true, a galaxy is classified as an active galaxy. This relation is an empirical refinement of a maximum starburst line by Kewley *et al.* (2001), above which star-forming galaxy models cannot produce examples. However, the intermediate region below the empirical

https://doi.org/10.1017/S1743921314010990 Published online by Cambridge University Press

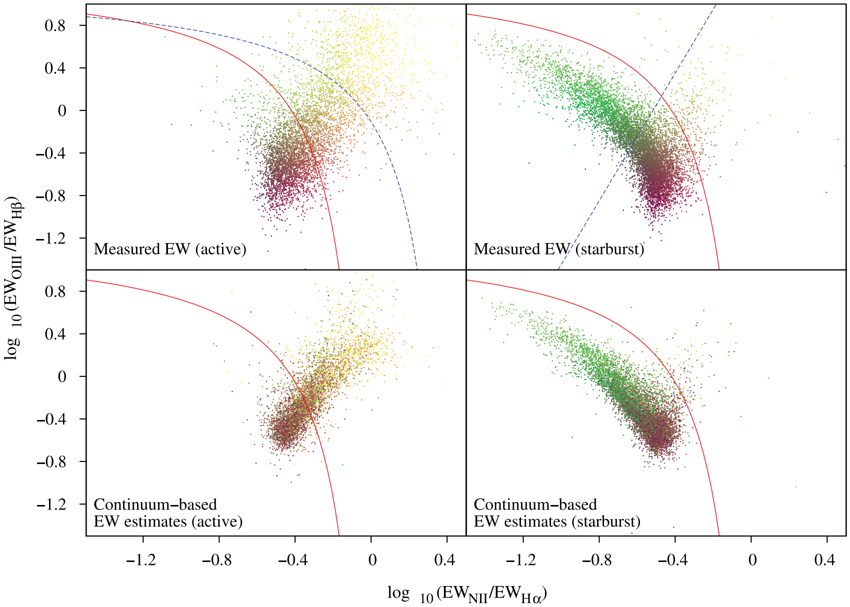

Figure 2. The BPT diagrams of active and star-forming galaxies, as classified by the support vector machine. It is observable that here the stellar continuum-based fitting projects galaxies to belong to the other class only rarely. The training sets for each class lie above the dashed lines on the top two images.

line – where the star-forming and active galaxy branches of the BPT diagram meet – has not yet been described in detail. The BPT diagrams of measured and fitted emission line equivalent widths for our sample are shown on Fig. 1.

4. Support vector machine classification

Based on the results shown on Fig. 1, we attempted to classify galaxies into the active and star-forming categories using the information contained in both their continua and emission lines. We performed principal component analysis on the logarithm of emission line equivalent widths, keeping the first four components. We selected by hand the relevant dimensions in the $5 + 4$ dimensional stellar continuum plus emission line PCA space, then trained a support vector machine on definitive examples. The results of the new classification are plotted on Fig. 2.

5. Acknowledgements

This research was supported by the OTKA-103244 grant of Hungary.

References

Baldwin, J. A., Phillips, M. M., & Terlevich, R. 1981, *PASP*, 93, 5

Kauffmann, G., Heckman, T. M., Tremonti, C., Brinchmann, J., Charlot, S., White, S. D. M.., Ridgway, S., Brinkmann, J., Fukugita, M., Hall, P., Ivezic, Z., Richards, G., & Schneider, D. 2003, *MNRAS*, 346, 1055

Kewley, L. J., Dopita, M. A., Sutherland, R. S., Heisler, C. A., & Trevena, J. 2001, *ApJ*, 556(1), 121

https://doi.org/10.1017/S1743921314010990 Published online by Cambridge University Press