MANIFOLD LEARNING IN THE CONTEXT OF QUASAR SPECTRAL DIVERSITY

I. JANKOV, D. ILIĆ and A. KOVAČEVIĆ

Department of Astronomy, Faculty of Mathematics, University of Belgrade,
Studentski trg 16, 11000 Belgrade, Serbia
E–mail: isidora_jankov@matf.bg.ac.rs

Abstract. Here we discuss the interpretation of quasar spectral diversity by subjecting quasars to non-linear treatment using a manifold learning technique called locally linear embedding (LLE). We apply the LLE analysis to a sample of type 1 quasars, for which the spectral features are taken from a Sloan Digital Sky Survey (SDSS) data catalog, which counts a total of ∼14,600 low-redshift objects. When we have a large number of parameters describing objects in question, LLE can be useful because it aims to find a low-dimensional representation of the original data set while preserving the geometry of the local neighborhoods within the data. We have shown that LLE can be used as a contextual tool in the search for essential correlations in the spectral parameters of SDSS quasars.

1. INTRODUCTION

Quasars are extremely bright and compact energy sources located in the central regions of some galaxies. Their extreme brightness is the result of the accretion of the material onto a supermassive black hole (Salpeter 1964, Shakura & Sunyaev 1973). This process creates a complex environment where the emission can be observed over a wide range of frequencies (Netzer 2013). The presence of different relationships between spectral parameters in different quasar populations originating from both physical and orientation effects results in a staggering diversity in their spectra (e.g., Netzer 2015).

In order to study this diversity, we need relatively large number of high quality spectra and appropriate methodology which can assist in finding the main trends in the data. One such method is the principal component analysis (PCA), which belongs to a family of linear dimensionality reduction methods. Identification of eigenvectors, or most dominant relationships, and their association with a fractional contribution to the data set's total variance is an advantageous PCA property. Many authors have utilized PCA to study quasar phenomenology, applying it to the measured spectral parameters (Boroson & Green 1992, Grupe 2004, Grimes et al. 2004, Hamilton et al. 2008, Jankov & Ilić 2020a) or directly on spectra (Francis et al. 1992, Brotherton et al. 1994, Shang et al. 2003, Yip et al. 2004, Ludwig et al. 2009).

One of the pioneering research papers that shed light on the complexity of the quasar parameter space was done by Boroson & Green (1992) where they applied

PCA to a sample of 87 low-redshift quasars from the Palomar–Green Bright Quasar Survey (Schmidt & Green 1983). They have identified an anti-correlation between peak flux of [O III] λ5007 Å and the equivalent width (EW) ratio of Fe II and broad Hβ ($R_{\rm FeII}$) as the main trend in their sample (the first principal component or Eigenvector 1). The full width at half maximum (FWHM) of the broad Hβ was also found to be well correlated with this component. The Eddington ratio was suspected as the driving mechanism behind this trend, but more convincing arguments were provided much later (Marziani et al. 2001, Boroson 2002, Shen & Ho 2014). Expanding further on the PCA analysis of Boroson & Green and also on correlations that emerged from ROSAT (e.g., Wang et al. 1996), Sulentic et al. (2000a) proposed an empirical formalism, analogous to the Hertzsprung-Russell diagram for stars, which allows contextualization of quasar spectral diversity for type 1 sources. They have found that quasars populate an elbow-shaped main sequence in the FWHM Hβ - $R_{\rm FeII}$ plane (E1 optical plane) driven by Eddington ratio convolved with the line of sight orientation. Using the main sequence, two main quasar populations (pop A and pop B) with significantly different physical properties are identified (e.g. Sulentic et al. 2000a,b, Sulentic et al. 2011, Marziani et al. 2018), as well as the highly accreting extreme pop A (Marziani & Sulentic 2014).

All above correlations in quasars were based on PCA results, which is a linear method, and its low dimensional projections cannot account for non-linear relationships that may emerge in the complex real-world data sets, such as quasar spectral data. Therefore, we select to utilize manifold learning (i.e., non-linear dimensionality reduction) and analyze correlations revealed by newly obtained projections. All manifold learning algorithms assume that real-world data sets lie on a low-dimensional manifold embedded in the high-dimensional space. The goal is to unravel the manifold and project it to a low-dimensional space while retaining the geometrical relationships between data points as much as possible. There are several manifold learning algorithms, but here we select to use locally linear embedding or LLE (Roweis & Saul 2000), since it preserves the geometry of the local neighbourhoods within the data, while staying computationally inexpensive and having only two free parameters. It also has a proven record of application in the astronomical context (e.g., Vanderplas & Connolly 2009, Daniel et al. 2011, Matijevič et al. 2012).

Here we give preliminary results of the analysis of the application of the LLE algorithm to a sample of low-redshift quasars. The paper is organized as follows: in Sec.2 we briefly describe the data and the LLE method, in Sec. 3 we give the first results, and in Sec.4 we outline our conclusions.

2. DATA AND METHOD

We use a sample of type 1 low-redshift (z < 0.35) quasars taken from the Sloan Digital Sky Survey Data Release 7 quasar catalog with measured optical spectral properties (Liu et al. 2019). The sample comprises of EW, FWHM and line luminosity (L) for broad lines (Hα^b and Hβ^b), EW and L for narrow lines (Hα^n, Hβ^n, [O III]λ5007, [O III]λ6300, [N II]λ6583, [S II]$\lambda\lambda 6716, 6731$), continuum luminosity at 5100Å (log L_{5100}), $R_{\rm FeII}$ and luminosity of Fe IIλ4570. During the object selection process, we only included objects with no null values for all previously mentioned spectral parameters. As a result, the sample size was reduced from ~14,600 down to 6236 objects. As LLE can be sensitive to outliers (Vanderplas & Connolly 2009), we made a choice to use density distribution in the E1 optical plane (FWHM Hβ - R_{FeII}) to locate and remove objects residing in the extremely low density regions. Although, we do have a total of 23 input parameters, we have chosen this particular plane for its proven high contribution to the total variance in quasar samples (Boroson & Green 1992, Sulentic et al. 2000a,b, Shen & Ho 2014). After the outlier removal, we were left with a final sample of 6023 objects. We need also to select the number of nearest neighbors (k) and the number of output dimensions (d) - two free parameters needed by the algorithm, which was done as described in Jankov et al. (2020b).

3. RESULTS AND DISCUSSION

We apply the LLE algorithm (setting $d = 3$ and $k_{opt} = 16$) to 6023 broad line quasars described by 23 spectral parameters using the scikit-learn Python library (Pedregosa et al. 2011), which results in three projections (Figure 1). We focus our attention to the $c_2 - c_3$ projection (*main projection*, further in the text), where the two of three identified branches are clearly distinguished (Figure 1 upper and lower left panel). Locations of objects in the LLE projections are determined by their proximity in the original 23-dimensional parameter space. Objects grouped in the projection have similar spectral parameters and can be potentially described as a distinct quasar population. Evidently, objects with different broad line structure, kinematics, and spectral characteristics described by pop A, B, and xA (e.g. Marziani et al. 2018) are clearly separated in two of three LLE projections, the main projection and $c_1 - c_3$ (see bottom panels in Figure 1). The axes (components) of the projections have no physical meaning (for details see Roweis & Saul, 2000). The main projection identifies object's progression along the N-branch (Narrow line branch) where many parameters associated with narrow emission lines vary. The N-branch is orthogonal to a branch identified as the quasar main sequence (defined as M-branch, Figure 1, bottom left). To study the correlations along this track, bins of the same sizes are created, and mean values of all input parameters are taken for each bin. Bins can be thought of as counterparts to composite spectra. The potentially outlying objects populating low density regions in the projection are eliminated (∼2% of the total sample).

Most prominent trends were spotted, namely the increase of narrow line EW (Figure 2) and narrow line luminosity along the branch, but also the decrease in continuum luminosity at 5100 Å and broad line luminosity (Figure 3). In addition to bin edges, the upper panel of Figure 1 shows the smooth increase in EW of [O III]λ5007, clearly indicating that the trend is indeed real. The N1 bin could contain objects with a range of different luminosities originating from another branch that hides behind our projection (Figure 3). This could produce higher average luminosity in the first bin as some objects would actually be far away in the projected space, and therefore have dissimilar properties in the obtained 3-dimensional LLE projection. That branch is identified as the L-branch (or luminosity branch) and it can be seen on the left side of $c_1 - c_2$ and $c_1 - c_3$ projections in Figure 1 (bottom middle and right panels). Additional analysis of the L-branch is needed to understand the relationships involving luminosities which is out of the scope of this contribution.

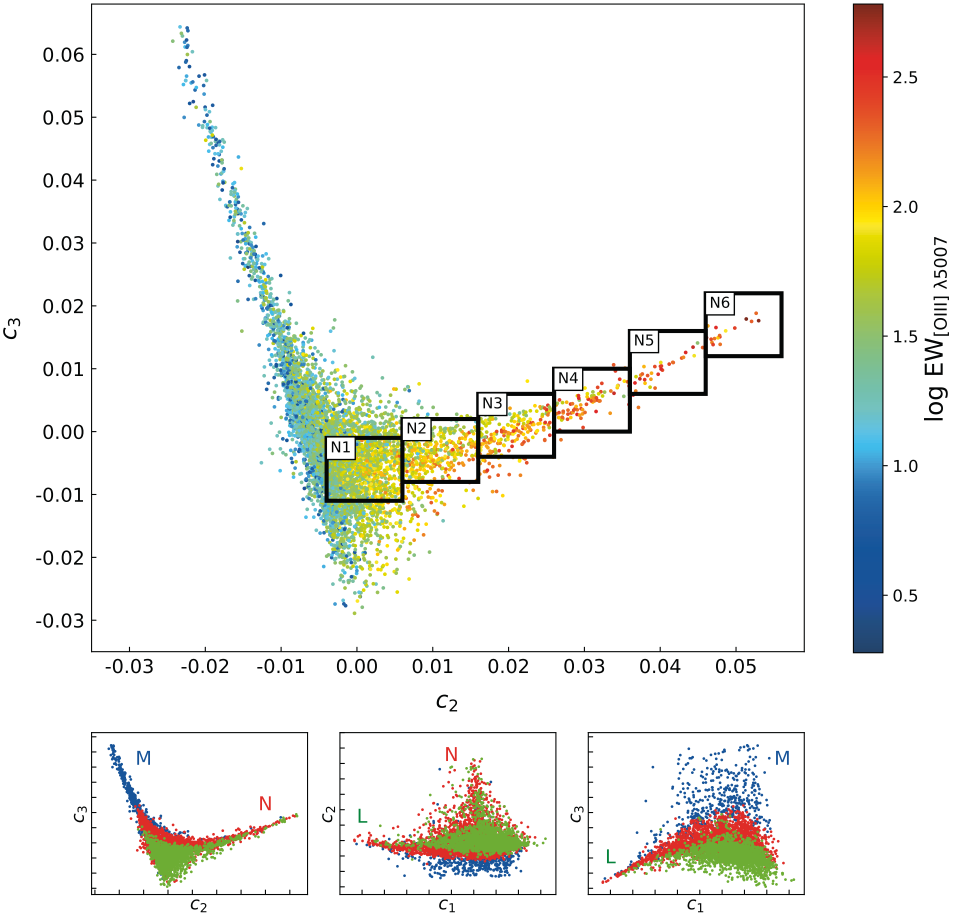

Figure 1: *Upper panel*: The main LLE projection. Number of dimensions is reduced from 23 down to three. This projection was chosen for analysis because of its clear distinction of objects that populate two different branches. The N-branch (Narrow line branch) is sampled by six equally sized bins. Axes are in arbitrary units and correspond to second and third component of LLE decomposition. The colormap indicates the gradient of log EW [O III] λ5007 in the projection. *Bottom panels*: All three LLE projections. Populations xA, A and B are marked with blue, red and green colors, respectively. Different branches are labeled by letters M (Main-sequence branch), N (Narrow line branch) and L (Luminosity branch).

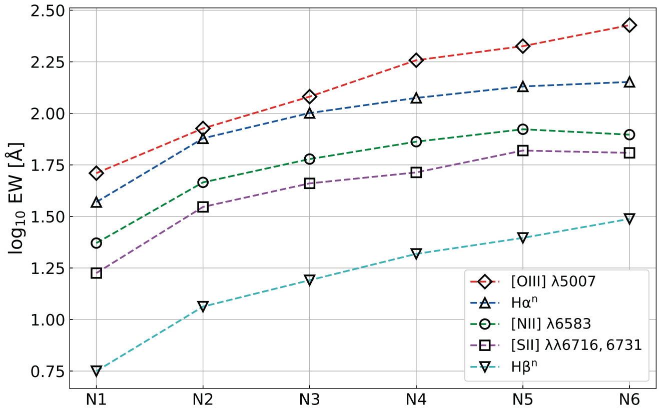

Figure 2: Progression of EW values of [O III] λ5007, narrow Hα, [N II] λ6583, [S II] $\lambda\lambda$6716, 6731 and narrow Hβ indicated by diamonds, triangles, circles, squares and reversed triangles, respectively. Symbols represent the mean values of log EW for each bin. Bins are ordered according to their location on the projection: starting from objects in bin N1 at the core of the projection and going towards the peak of the branch populated by objects from bin N6. Symbols are connected by dashed lines of different colors for clarity.

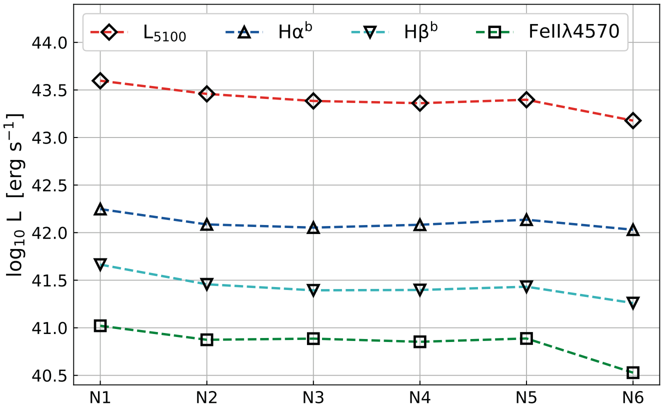

Figure 3: Progression of luminosity of the continuum at 5100 Å (L_{5100}), broad line luminosity (Hα and Hβ) and the luminosity of FeII λ4570. Bins are ordered as described in Figure 2.

4. CONCLUSIONS

We have applied the LLE algorithm to a large sample from SDSS catalogue of quasar spectral properties in order to study the quasar high-dimensional parameter space (23 spectral parameters). We have found a low-dimensional projection which clearly identifies three branches, and we discuss in more details the properties of the objects along the branch of increasing narrow emission lines, i.e., the N-branch. The analysis of bins along the N-branch shows the dominant trend producing the branch itself - the variation of EW of all narrow lines from our sample indicating the presence of a separate quasar population characterized by apparently strong narrow emission lines (e.g., Ludwig et al. 2009). The continuum luminosity appears to decrease along the branch, but this effect must be studied in tandem with other branches because objects from those branches may contaminate the first bin. We conclude that manifold learning techniques, LLE in particular, have proven to be immensely valuable tools for the deep understanding of complex data sets such as quasar spectral data and it may be proven even more useful if applied to ever more complex astronomical data sets of the future.

References

Boroson, T. A.: 2002, *The Astrophysical Journal*, **565**, 78.
Boroson, T. A., Green, R. F.: 1992, *Astrophysical Journal Supplement*, **80**, 109.
Brotherton, M. S. et al.: 1994, *The Astrophysical Journal*, **430**, 495.
Daniel, S. F. et al.: 2011, *The Astronomical Journal*, **142**, 203.
Francis, P. J. et al.: 1992, *The Astrophysical Journal*, **398**, 476.
Grimes, J. A., Rawlings, S., Willott, C. J.: 2004, *MNRAS*, **349**, 503.
Grupe, D.: 2004, *The Astronomical Journal*, **127**, 1799.
Hamilton, T. S., Casertano, S., Turnshek, D. A.: 2008, *The Astrophysical Journal*, **678**, 22.
Jankov, I., Ilić, D.: 2020a, *Contrib. Astron. Obs. Skaln. Pleso*, **50**, 350.
Jankov, I., Ilić, D., Kovačević, A.: 2020b, *Publ. Astron. Obs. Belgrade*, **99**, 291-294.
Liu, H.-Y. et al.: 2019, *Astrophysical Journal Supplement*, **243**, 21.
Ludwig, R. R. et al.: 2009, *The Astrophysical Journal*, **706**, 995–1007.
Marziani, P. et al.: 2001, *The Astrophysical Journal*, **558**, 553.
Marziani, P. et al.: 2018, *Frontiers in Astronomy and Space Sciences*, **5**, 6.
Marziani, P., Sulentic, J.: 2014, *MNRAS*, **442**, 1211-1229.
Matijevič, G. et al.: 2012, *The Astronomical Journal*, **143**, 123.
Netzer, H.: 2013, The Physics and Evolution of Active Galactic Nuclei, by Hagai Netzer, Cambridge, UK: Cambridge University Press, 2013.
Netzer, H.: 2015, *Annu. Rev. Astron. Astrophys.*, **53**, 365-408.
Pedregosa, F. et al.: 2011, *Journal of Machine Learning Research*, **12**, 2825-2830.
Roweis, S. T., Saul, L. K.: 2000, *Science*, **290**, 2323.
Salpeter, E. E.: 1964, *The Astrophysical Journal*, **140**, 796.
Schmidt, M., Green, R. F.: 1983, *The Astrophysical Journal*, **269**, 352-374.
Shakura, N. I., Sunyaev, R. A.: 1973, *Astronomy and Astrophysics*, **500**, 33.
Shang, Z. et al.: 2003, *The Astrophysical Journal*, **586**, 52-71.
Shen, Y., Ho, L. C.: 2014, *Nature*, **513**, 210.
Sulentic, J. W., Marziani, P., Dultzin-Hacyan, D.: 2000a, *Annu. Rev. Astron. Astrophys.*, **38**, 521.
Sulentic, J. W. et al.: 2000b, *The Astrophysical Journal*, **536**, L5-L9.
Sulentic, J. W., Marziani, P., Zamfir, S.: 2011, *Baltic Astronomy*, **20**, 427-434.
Vanderplas, J., Connolly, A.: 2009, *The Astronomical Journal*, **138**, 1365-1379.
Wang, T., Brinkmann, W., Bergeron, J.: 1996, *Astronomy and Astrophysics*, **309**, 81.
Yip, C. W. et al.: 2004, *The Astronomical Journal*, **128**, 2603-2630.