Publ. Astron. Obs. Belgrade No. 99 (2020), 291 - 294 *Contributed Paper*

APPLICATIONS OF MANIFOLD LEARNING TECHIQUES TO SPECTRAL CLASSIFICATION OF QUASARS

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, dilic@matf.bg.ac.rs, andjelka@matf.bg.ac.rs

Abstract. In the last three decades, application of different techniques, such as correlation matrix analysis and principal component analysis (PCA), on different spectral parameters of type 1 quasars, has revealed that they occupy a specific parameter space, analogous to main sequence of stars revealed by H-R diagram. Here we investigate a sample of low-redshift ($z < 0.39$) type 1 quasars described by ten spectral features taken from a Sloan Digital Sky Survey catalog using a manifold learning technique called locally linear embedding (LLE). Preliminary results of our investigation indicate that LLE performs better than PCA in terms of clearer visual and functional representation of the quasar main sequence, which may help in future large survey classification of type 1 quasars based on their spectral properties.

1. INTRODUCTION

Much progress was made in the field of active galaxies when larger quasar samples became available, providing spectra with high S/N (Schmidt & Green, 1983). In one of the pioneering research papers that made use of these observations, principal component analysis (PCA) was applied to a sample of 87 quasars (Boroson & Green, 1992). The PCA revealed the main trend in the data – an anti-correlation between [O III] λ5007 Å equivalent width (EW) and the EW ratio of Fe II and broad Hβ - R_{FeII} described by the first principal component. Full width at half maximum (FWHM) of broad Hβ line was also found to be associated with this component (hence the name "Eigenvector 1" or E1). These results set course to many other investigations of spectral properties of quasars based on PCA (Marziani, et al., 2001; Shang, et al., 2003; Grupe, 2004; Yip, et al., 2004; Wang, et al., 2006; Zamfir, et al., 2008; Kuraszkiewicz, et al., 2009) and present the basis for definition of 4DE1 (Sulentic, et al., 2000). More notable example of important leap in understanding the quasar properties was the identification of two main quasar populations (A and B) with different broad line structure, kinematics and spectral

characteristics in optical, UV and X-ray wavelengths (Sulentic, et al., 2000) and the recognition of the Eddington ratio convolved with the line-of-sight orientation of the source as the potential driving mechanism behind the quasar main sequence (MS) revealed by E1 (Marziani, et al., 2001; Shen & Ho, 2014).

In the era of large astronomical surveys, astronomers deal with data sets with ever increasing number of dimensions. Dimensionality reduction (DR) techniques can help us decide which parameters or combinations thereof carry the most information in the data set. In addition, these techniques can be a powerful tool to visualize high-dimensional data while retaining large portion of information contained in the data and potentially help in classification (for a review of DR methods see Łukasik et al., 2016 and Baron, 2019). It was repeatedly demonstrated that PCA is highly effective in tasks of finding important linear relationships in high-dimensional data, but it has its shortcomings when applied to inherently non-linear data sets, such is the case with galaxy spectra where variations of some spectral parameters can be non-linear functions of the galaxy type. This problem can be alleviated by non-linear DR techniques (i.e. manifold learning) such as locally linear embedding – LLE. Goal of LLE is to find a low-dimensional representation of original data set while preserving the geometry of local neighborhoods within the data (Roweis & Saul, 2000). Motivated by previous applications of LLE in astronomical context (e.g. Vanderplas & Connolly, 2009; Daniel et al., 2011; Matijevič et al., 2012) our goal was to see how the interpretation of quasar spectral diversity can be improved by non-linear treatment of type 1 quasars in the context of E1 parameters.

2. DATA ANALYSIS

We use measured spectral properties from the Sloan Digital Sky Survey Data Release 7 quasar catalog (Shen, et al., 2011) and define our sample so it contains only low-redshift objects ($z < 0.39$) with measured both narrow and broad Hα and Hβ components, as well as [O III] λ5007 Å emission line, continuum luminosity ($\log L_{5100}$) and restframe EW of Fe within 4435 – 4685 Å used for calculation of R_{FeII}. Continuum luminosity was corrected for the host galaxy starlight contamination using the empirical fitting formula given by Shen et al. (2011) in their Eq. 1. LLE can be very sensitive to outliers, so before applying the algorithm to our sample, we looked at the density distribution of quasars in E1 optical plane (FWHM Hβ - R_{FeII}) and removed points residing in extremely low-density regions. After excluding outliers from the optical plane using this simple method, we were left with 3720 objects in our sample.

One of the strengths of LLE is the fact that it needs only one free parameter – the number of nearest neighbors (k). The algorithm uses this parameter to learn the local geometry of the manifold. If the value of k is too low, the manifold could be falsely divided into disjoint sub-manifolds. In contrast, if k is too high, the local neighborhood no longer lies on approximately linear surface and Euclidian distances are no longer valid metric for finding the nearest neighbors, leading to

false interpretation of the manifold. Having this in mind, one needs to be cautious when deciding the value of k. Approach that we have used was similar to one used by Tenenbaum et al. (2000) and Kouropteva et al. (2002). In order to choose a value for k, we have calculated a matrix of pairwise geodesic distances of original parameter space, as well as for the resulting low-dimensional space. Then, we compared the matrices by calculating the modified RV coefficient (Smilde, et al., 2008), which is a measure of correlation between two matrices. The process is repeated for a range of values of k, in our case $4 \leq k \leq 30$. Following this approach, we have obtained an optimal number of nearest neighbors ($k_{\mathrm{opt}} = 12$), the one with highest value of modified RV coefficient.

3. RESULTS AND DISCUSSION

After initial preparation of the sample, we have applied LLE with $k = 12$ and specified the output dimension to three in order to provide a visual representation of the original ten-dimensional parameter space. Ten spectral parameters that were used as input in LLE are: broad Hα and Hβ (their EW, FWHM, line luminosity - L), [O III] λ5007 Å (their EW and L), continuum luminosity (log L_{5100}) and R_{FeII}.

Left panel of Fig. 1 presents the resulting 3D space where different quasar populations occupy distinct regions, emphasizing their spectral differences. Objects belonging to populations xA (highly accreting extension of population A, Marzinai & Sulentic, 2014) and B present two extremes of the MS following the direction indicated with an arrow, while population A appears to be an intermediate class. Contrary to PCA, the interpratation of the LLE projection is found in the relationship between the points rather than from the axes (components). In our case, axes present three eigenvectors with lowest nonzero eigenvalues that are obtained by solving a sparse matrix eigenvalue problem which is the last step of the algorithm when low-dimensional projection with perserved local neighborhoods is found (for details see Roweis & Saul, 2000). The resulting LLE projection illustrates the relationships between spectral parameters closest to the original relationships in high-dimensional space and we can use this projection to follow variations of different physical parameters, as presented on the right panel of Fig. 1 in the case of Eddington ratio, the probable principal driving mechanism of the quasar MS. In this way, we can identify parameters that have high correlation with the MS which is an important information needed to further investigate the nature of different quasar populations and potentially help in future large survey classification of type 1 quasars based on their spectral properties.

Preliminary results of our investigation indicate that LLE can be a powerful tool in data exploration and identification of objects with distinct spectral properties. The analysis of our data set showed that it is possible to find three eigenvectors, giving projection in 3D that contains additional information potentially lost in the 2D projection. Our results have confirmed the presence of the quasar MS driven by Eddington ratio in a space with maximum preservation of the original manifold geometry. Further research is needed to investigate the nature of the quasar

populations, mainly by including more spectral parameters as input for LLE and by applying the algorithm to more recent spectral data.

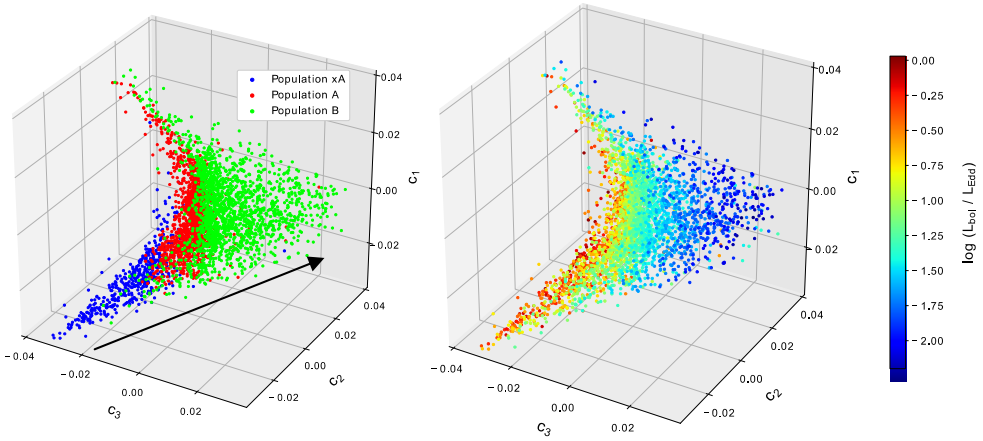

Figure 1: 3D projection of the original manifold embedded in ten-dimensional space. Axes are in arbitrary units and correspond to three components of LLE decomposition. Left – populations xA, A and B are marked with blue, red and green, respectively. The black arrow points in the direction of the quasar MS. Right – gradient of the Eddington ratio in the resulting projection. Direction of the gradient is close to the direction of the MS.

References

Baron, D. : 2019, *arXiv e-prints*, arXiv:1904.07248
Boroson, T. A., Green, R. F. : 1992, *Astrophysical Journal Supplement,* **80**, 109.
Daniel, S. F. et al. : 2011, *The Astronomical Journal,* **142**, 203.
Grupe, D. : 2004, *The Astronomical Journal*, **127**, 1799-1810.
Kouropteva et al. : 2002, *First International Conference of Fuzzy Systems,* 359-363.
Kuraszkiewicz, J. et al. : 2009, *The Astrophysical Journal*, **692**, 1180-1189.
Łukasik, S. et al. : 2016, *Open Physics*, **14**, 64.
Marziani, P. et al. : 2001, *The Astrophysical Journal*, **558**, 553-560.
Marziani, P., Sulentic, J. W. : 2014, *MNRAS,* **442**, 1211-1229.
Matijevič, G. et al. : 2012, *The Astronomical Journal*, **143**, 123.
Roweis, S. T., Saul, L. K. : 2000, *Science*, **290**, 2323-2326.
Schmidt, M., Green, R. F. : 1983, *The Astrophysical Journal*, **269**, 352-374.
Shang, Z. et al. : 2003, *The Astrophysical Journal,* **586**, 52-71.
Shen, Y., Ho, L. C. : 2014, *Nature,* **513**, 210-213.
Shen, Y. et al. : 2011, *The Astrophysical Journal Supplement,* **194**, 45.
Smilde, A. K. et al. : 2008, *Bioinformatics,* **25**, 401-405.
Sulentic, J. W. et al. : 2000, *Annual Review of Astronomy and Astrophysics,* **38**, 521-571.
Tenenbaum, J. B. et al. : 2000, *Science*, **290**, 2319-2323.
Vanderplas, J., Connolly, A. : 2009, *The Astronomical Journal,* **138**, 1365-1379.
Wang, J. et al. : 2006, *The Astrophysical Journal,* **638**, 106-119.
Yip, C. W. et al. : 2004, *The Astronomical Journal,* **128**, 2603-2630.
Zamfir, S. et al. : 2008, *MNRAS,* **387**, 856-870.