University of Washington
Survey Science Group

Finding the Odd One Out in Large Spectral Surveys

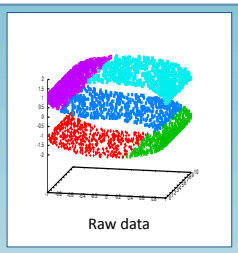
Scott F. Daniel, Andrew J. Connolly, Jake Vanderplas (University of Washington), Jeff Schneider, Liang Xiong (Carnegie Mellon University)

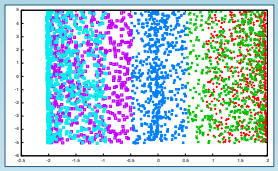
Through shear volume of data, next generation surveys will provide us with unprecedented and detailed information about the full distribution of astronomical sources. Anomalous events influenced by the the most extreme and interesting physics will no longer be relegated to the dustbin of "small number statistics." Unfortunately, that same volume of data will render the task of culling these extreme events from the background of ordinary stars and galaxies virtually impossible. Both the number of events and the dimensionality of the data (e.g. a spectral energy distribution measured in 4000 wavelength bins) exist well outside the reasonable limits of human processing In this context, we seek algorithms to project N>>1 dimensional data down to 2 or 3 effective dimensions, preserving the physics of the correlations within the unprojected data. Inspection in these effective dimensions then allows us to identify both objects that resemble one another (classification of objects) and objects that resemble nothing at all (anomaly detection). We consider Local Linear Embedding, which attempts to reconstruct data points only from their nearest neighbors, preserving the non-linear relationships between different neighborhoods. We use spectra from the SDSS to show how this technique can aid in the classification of astronomical sources.

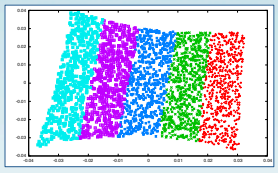
Local Linear Embedding (LLE) [1] reconstructs each of the original N-dimensional data points as a linear combination of its nearest neighbors, i.e.

$$\vec{x}^{(i)} = \sum_j w_{ij}\vec{x}^{(j)}$$

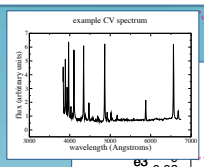
where w is a weight matrix and j sums over the nearest neighbors. When projecting down to the reduced dimensionality, LLE preserves the non-linear relationships between the original data points by finding d-dimensional (d<<N) points y that are related by the same weight matrix w.

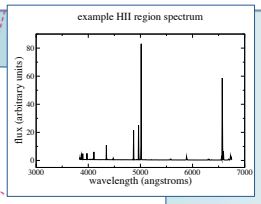


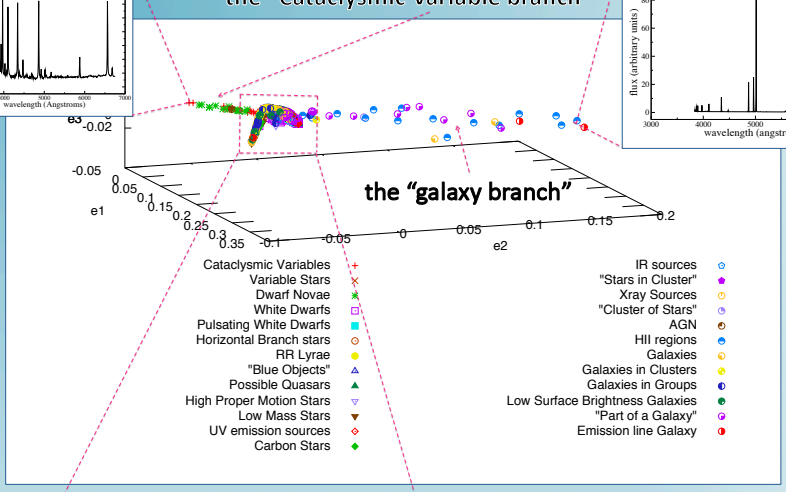


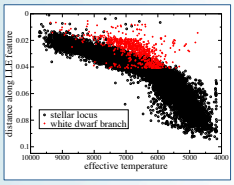


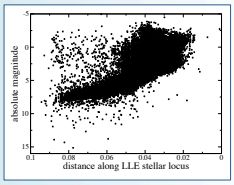
Above is an example of an LLE projection from highly non-linear three-dimensional data to its underlying linear two-dimensional counterpart. As a linear decomposition, Principal Component Analysis (PCA) does not correctly unwind the data.

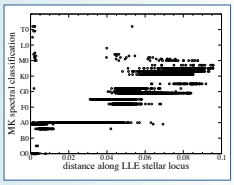


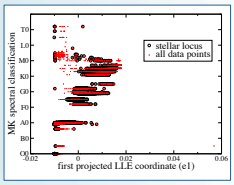






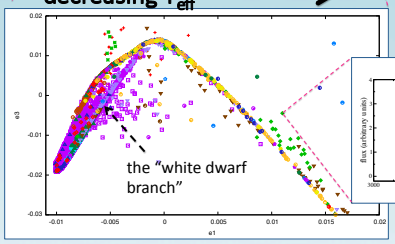


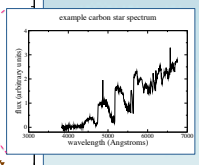




These plots show how position in LLE space correlates with a star's physical attributes (as determined by the sppParams pipeline [4-6]). The black points are data points that are located on the stellar locus. There seems to be a one-dimensional correlation with temperature, absolute magnitude, or spectral type, just as in the HR diagram.

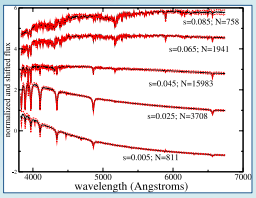
decreasing T_{eff} →





the "stellar locus"

We use LLE to project 49,529 SDSS DR7 [2] spectra down from a 500-dimensional wavelength bin space to a three-dimensional LLE space. We find that similar objects tend to cluster together in the LLE space. Object types are taken from the SIMBAD database [3]. Most stellar objects gather in the one-dimensional stellar locus feature plotted in detail above. Position in the stellar locus correlates with effective temperature, as shown in the figures on the right side of this poster.



The mean spectrum as a function of distance (s) along the stellar locus measured from the left corner. Red dashed curves show 1-σ variation about the mean. N is the number of spectra averaged.

References
[1] Roweis, S.T. and Saul, L.K. 2000, Science **290**, 2323
[2] Abazajian, K.N. *et al.* ApJ. S **182**, 543
[3] http://simbad.u-strasbg.fr/simbad/
[4] http://cas.sdss.org/astrodr7/en
[5] Lee, Y.S. *et al.* 2008 AJ **136**, 2022
[6] Lee, Y.S. *et al.* 2008 AJ **136,** 2050

We gratefully acknowledge support from the DOE Applied Mathematics Program through grant DE-FG02-87ER40315 and the NSF through grant IIS-0844580