Statistical Challenges in 21st Century Cosmology
Proceedings IAU Symposium No. 306, 2014
A. F. Heavens, J.-L. Starck & A. Krone-Martins, eds.

doi:10.1017/S1743921314013581

Darth Fader: Analysing galaxy spectra at low signal-to-noise

Adrienne Leonard*[1,2], Daniel P. Machado[2], Filipe B. Abdalla[1] and Jean-Luc Starck[2]

[1]Department of Physics and Astronomy, University College London
Gower Place, London WC1E 6BT, United Kingdom

[2]Laboratoire AIM, UMR CEA-CNRS-Paris 7, IRFU, Service d'Astrophysique
CEA Saclay, F-91191 Gif-sur-Yvette CEDEX, France

*email: adrienne.leonard@ucl.ac.uk

Abstract. Spectroscopic redshift surveys are an incredibly valuable tool in cosmology, allowing us to trace the distribution of galaxies as a function of distance and, thus, trace the evolution of structure formation in the Universe. However, estimating the redshifts from spectra with low signal-to-noise is difficult, and such data are often either discarded or require human classification of spectral lines to obtain the galaxy redshift. Darth Fader offers an automated method for estimating the redshifts of galaxies in the low signal-to-noise regime. Using a sophisticated, wavelet-based technique, galaxy spectra can be separated into continuum, line and noise components, and the lines can then be cross-correlated with template spectra in order to estimate the redshifts. Cross-matching of the identified lines then allows for a cleaning of the resulting catalogue, effectively removing the vast majority of erroneous redshift estimates and resulting in a highly pure, highly accurate redshift catalogue. Darth Fader allows us to effectively use low signal-to-noise galaxy spectra, and dramatically reduces the number of human hours required to do this, allowing spectroscopic surveys to probe deeper into the formation history of the Universe.

Keywords. galaxies: distances and redshifts, methods: data analysis, techniques: spectroscopic surveys

1. Introduction

Automated spectroscopic redshift estimation is typically carried out using either modelling, template matching, or cross-correlation techniques. In Darth Fader (Machado *et al.* 2013), we employ a cross-correlation method (see, e.g., Glazebrook *et al.* 1998). We assume that any galaxy spectrum can be represented as a linear combination of template spectra:

$$S_\lambda = \sum_i a_i T_{i\lambda} \tag{1.1}$$

where the index λ runs over the wavelengths sampled by the spectrograph, S_λ is the true spectrum of the galaxy and $T_{i\lambda}$ is a representative set of template spectra.

If the templates and galaxy spectra are binned on a logarithmic wavelength axis, a shift along this axis is directly proportional to $\log(1+z)$. The redshift can then be obtained by considering the cross-correlation between the galaxy spectra and the templates. However, cross-correlation methods require the templates and galaxy spectra to be continuum free. Templates can be obtained either from simulations or high signal-to-noise data. We use a principal component analysis to reduce the dimensionality of the problem, and to extract the important features of the template spectra.

https://doi.org/10.1017/S1743921314013581 Published online by Cambridge University Press

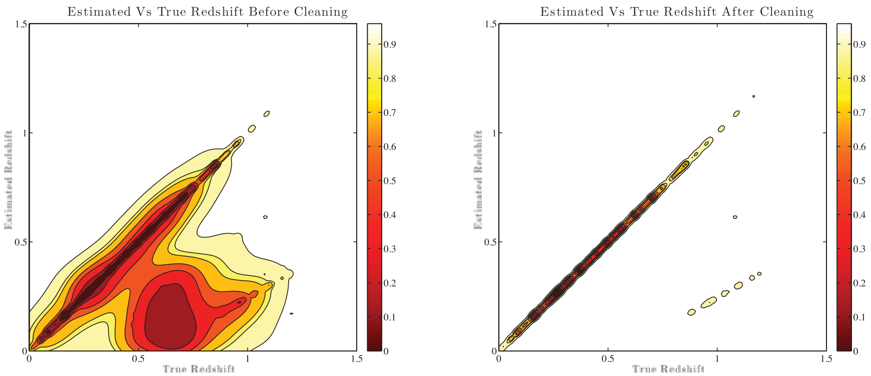

Figure 1. The distribution of estimated vs. true redshift for simulated galaxy spectra at a signal-to-noise of 2 in the r-band. The left panel shows the distribution before cleaning, where we obtain a success rate of 65.5%, while the right panel shows the distribution after cleaning, where we have retained 76.2% of the galaxy spectra and reached a success rate of 94.9%.

Darth Fader works by exploiting the sparsity properties of spectra in order to separate the line signal, continuum and noise from each spectrum in a blind, nonparametric way. Spectra are decomposed in a wavelet basis; the continuum is represented by the largest wavelet scale, and the noise is removed using an iterative procedure that identifies the most significant wavelet coefficients (those containing the sparse signal) using a False Discovery Rate method (Benjamini & Hochberg, 1995).

Thus Darth Fader is able to separate a spectrum into three components: lines, continuum and noise. The line spectra are then used to estimate the redshifts of the galaxies by cross-correlation. At high signal-to-noise, cross-correlation will yield a highly pure sample of redshift estimates. However, this method begins to fail in certain cases when the signal-to-noise of the galaxy spectrum is low. In this case, Darth Fader employs a cleaning step, whereby galaxies whose line spectra show very few features, or whose blueshifted features do not match any expected or prevalent lines, are excluded from further analysis.

2. Darth Fader Performance

To test the Darth Fader algorithm, we first generated a sample of simulated galaxy spectra (Jouvel *et al.* 2009) at a signal to noise of 2 in the r-band†. In Figure 1 we show the distribution of estimated vs. true redshift after cross-correlation with the templates. We obtained a correct redshift estimate for around 65.5% of the galaxies in the sample. The figure also shows the distribution of redshift estimates after cleaning. To clean the catalogue, we selected galaxies with 6 or more line features. Thus, we retained 76.2% of the sample and attained a success rate of 94.9% for this subsample.

We also applied Darth Fader to spectroscopic data from the WiggleZ survey (Drinkwater *et al.* 2010, 2014), which contains 225,415 low signal-to-noise spectra. The redshifts for all the galaxies in this survey were obtained by eye. Figure 2 shows the results of application of Darth Fader to a subsample of 1000 randomly-selected WiggleZ galaxies. Before cleaning, using cross-correlation we correctly matched redshifts for only $\sim$ 50% of the galaxies. To clean the catalogue, we selected galaxies whose rest-frame spectra showed features consistent with OII or Hα. We then matched 88% of the redshift estimates, retaining around $\sim$ 60% of the galaxies.This demonstrates the effectiveness of Darth Fader

† See Machado *et al.* (2013) for additional experiments with simulated data.

https://doi.org/10.1017/S1743921314013581 Published online by Cambridge University Press

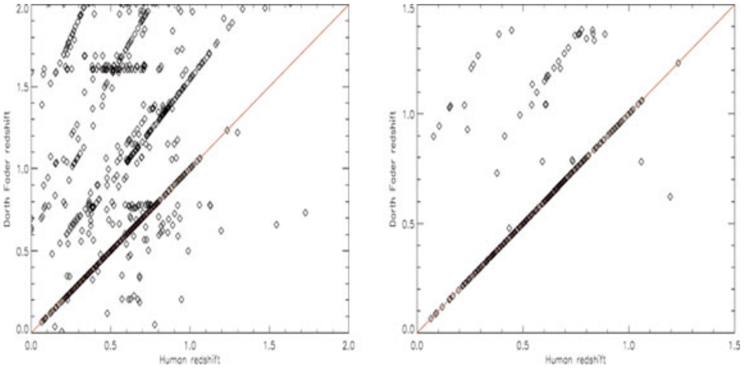

Figure 2. Darth Fader redshift (y-axis) vs Human redshift (x-axis) for a subsample of 1000 WiggleZ galaxy spectra. The left panel shows the distribution before cleaning, for which we obtain matches for around 50% of the galaxies, while the right panel shows the distribution after cleaning, where we retain around 60% of the galaxies with a redshift match rate of 88%.

to obtain accurate redshift estimates in the low-signal-to-noise regime with minimal data loss, and could represent a dramatic reduction in the number of human hours required to analyse such data, and a valuable cross-check on human-identified redshift estimates.

3. Summary

Darth Fader is a powerful tool for the improvement of redshift estimation without any a priori knowledge of galaxy composition, type or morphology. Our algorithm allows us to successfully estimate the continuum without the need for detailed modelling of the galaxy spectra, and to confidently make use of data at signal-to-noise levels that were previously beyond the reach of other techniques. This is demonstrated through extensive experiments using simulated data in Machado *et al.* (2013).

We have successfully applied this technique to real data from the WiggleZ survey (Leonard *et al.* 2014). Though this research is still in a preliminary stage, we are already able to correctly estimate the redshifts for around half of the galaxies in the survey in an automated and therefore fast way. Moreover, we can effectively separate the galaxies for which we expect to obtain a reliable redshift estimate, generating a high-purity subsample of galaxy redshift estimates. This is useful both for verification of human redshift estimates and to reduce the number of human hours required to analyse a survey such as WiggleZ. The Darth Fader software is publicly available at `http://cosmostat.org/darth_fader.html`

References

Benjamini, Y. & Hochberg, Y. 1995, *Journal of the Royal Statistical Society B*, 57, 289
Drinkwater, M. J. *et al.* 2010, *MNRAS*, 401, 1429
Drinkwater, M. J. *et al.* 2014, *in prep.*
Glazebrook, K., Offer, A. R., & Deeley, K. 1998, *ApJ*, 492, 98
Jouvel, S. *et al.* 2009, *A&A*, 504, 359
Leonard, A. *et al.* 2014, *in prep.*
Machado, D. P., Leonard, A., Starck, J.-L., Abdalla, F. B., & Jouvel, S. 2014, *A&A*, 560, 83

https://doi.org/10.1017/S1743921314013581 Published online by Cambridge University Press