DARTH FADER: Using wavelets to obtain accurate redshifts of spectra at very low signal-to-noise

D. P. Machado[1] A. Leonard[1] J.-L. Starck[1] F. Abdalla[2]

[1] CEA Saclay, IRFU, Service d'Astrophysique, Bt 709 - Orme des Merisiers, 91191 Gif-Sur-Yvette CEDEX, France.

[2] University College London, Department of Physics & Astronomy, Kathleen Lonsdale Building, Gower Place, London, WC1E 6BT, United Kingdom.

Introduction

Accurate determination of galactic redshift comes from the identification of key lines in their spectra. When these spectra are too noisy, they are often thrown away by cutting in flux/magnitude or signal-to-noise. We present DARTH FADER (**D**enoised and **A**utomatic **R**edshifts **Th**resholded with a **Fa**lse **De**tection **R**ate), which is a wavelet-based method for estimating galactic redshift, and empirically isolating the continuum of a spectrum. We employ a wavelet denoising on the spectrum using a *false detection rate* threshold, and counting the number of peaks in the cleaned spectrum. We choose cleaned spectra with three peaks to be good candidates since the presence of only two lines leads to an ambiguity in the estimated redshift.

Mock Catalogues

- Mock catalogues were generated using LePhare simulation program (Arnouts, et al. 1999) with 3000 top-hat filters for the template set, run in batches of 51 filters, with the first filter being the same reference filter in each batch, and CWW-Kinney templates (Coleman, et al. 1980; Kinney, et al. 1996) used throughout.
- The filters evenly span the $\log_{10}$ of the wavelength axis from 3.0 to 4.5, corresponding to a wavelength range of 1000Å to 31623Å. The galaxy catalogue was made in a similar manner, but using only 2000 top-hat filters, resulting in a wavelength range of 3160Å to 31623Å (3.5 to 4.5 on the log axis).
- The template set consisted of 259 noiseless galaxies, and the galaxy catalogue of 3775.
- Multiple galaxy catalogues were constructed from this base set, by adding wavelength independent Gaussian noise on the spectrum. Multiple fixed signal-to-noise, and a single uniformly mixed signal-to-noise, catalogues were constructed spanning the range 0.1 to 3.0.

Cross-Correlation and PCA

- Darth Fader utilises a standard PCA and cross-correlation procedure, similar to that of Glazebrook, et al. (1998).
- It is possible to construct a set of orthonormal eigentemplates, **E**, from any template set, **T**, via a PCA procedure:

$$E_{j\lambda} = \frac{\sum_i R_{ij}^T T_{i\lambda}}{\sqrt{\Lambda_j}},$$

  where **R** represents the matrix of eigenvectors that have been descendingly ordered via their corresponding eigenvalues Λ, in turn obtained from diagonalising the correlation matrix of **T**.
- The estimate of the goodness-of-fit to galaxy spectrum, G_λ, can be found by computing the minimum distance via a standard χ^2 :

$$\chi^2(\Delta) = \sum_\lambda \frac{w_\lambda^2}{\sigma_\lambda^2}\Big[G_\lambda - \sum_i a_i(\Delta)\, T_{i(\lambda+\Delta)}\Big]^2,$$

  where a_i are expansion coefficients that can be determined from minimisation, and Δ is the linear translation along the log-wavelength axis undergone by the template. Provided we have chosen to bin the spectra logarithmically, then $\Delta = \log(1+z)$. For our method we set the weighting function, w_λ^2, and the variance, σ_λ^2, as wavelength independent and constant.
- By substituting the template set, **T**, for the eigentemplate set, **E**, using the orthogonality condition between eigentemplates and the properties of the convolution theorem, we can significantly simplify this to finding the *maximum* of the following:

$$\widetilde{\chi}^2(\Delta) = \sum_{i=1}^{N}\Big[\mathcal{F}^{-1}(\hat{G}_k\,\hat{E}_{ik})\Big]^2,$$

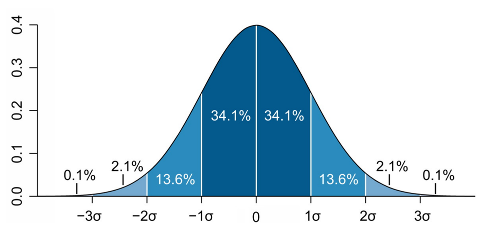
  where $\hat{}_k$ and $\mathcal{F}^{-1}$ represent Fourier and inverse Fourier transforms respectively.
- The $\widetilde{\chi}^2$ function reaches a maximum (and thus the χ^2, a minimum) when the shift of the templates along the log-wavelength axis corresponds to the true shift of the galaxy spectrum, so that the redshift is estimated to be where $\Delta = \Delta_{\widetilde{\chi}} = \Delta|_{\widetilde{\chi}=\widetilde{\chi}_{max}}$; giving:

$$z_{est} = 10^{\delta_s \Delta_{\widetilde{\chi}}} - 1,$$

  where δ_s is the grid spacing on the log-wavelength axis.

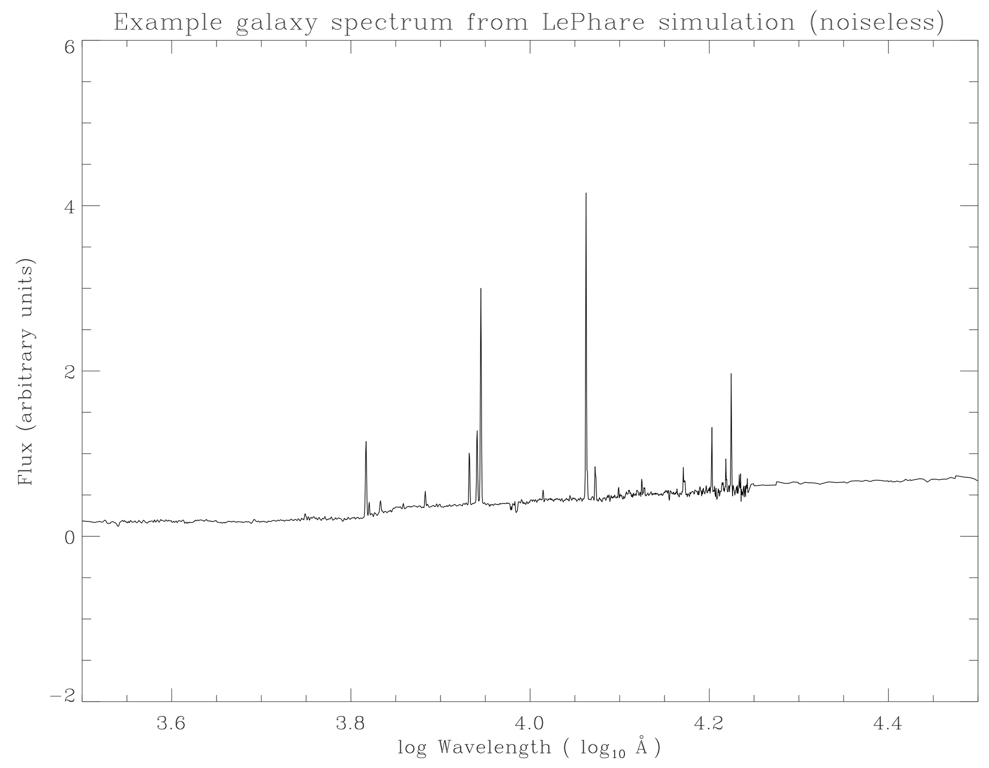
Wavelets

- Darth Fader uses an isotropic wavelet transform, and wavelet denoising with a false detection rate (FDR) threshold. We use the *star1d* and *mr1d_filter* algorithms developed by Starck, et al. (1994)*.
- A wavelet transform yields a representation of a signal on different scales (much like city maps).
- It is possible to edit these scales individually, before reconstituting them; this process is known as *wavelet filtering*.
- This is useful for extracting signal information from noisy data, since noise features typically are on smaller scales.
- FDR thresholding is way of defining signal relative to noise, such that the resultant denoising contains a very small percentage of false detections. It is more sophisticated than, for example, a simple n-σ clipping.
- The FDR threshold alpha, which is the maximum allowed fraction of false detections, is specified in mr1d_filter relative to an n-σ of a Gaussian PDF, such that specifying 2σ in mr1d_filter results in an FDR alpha of 4.55% false detections.



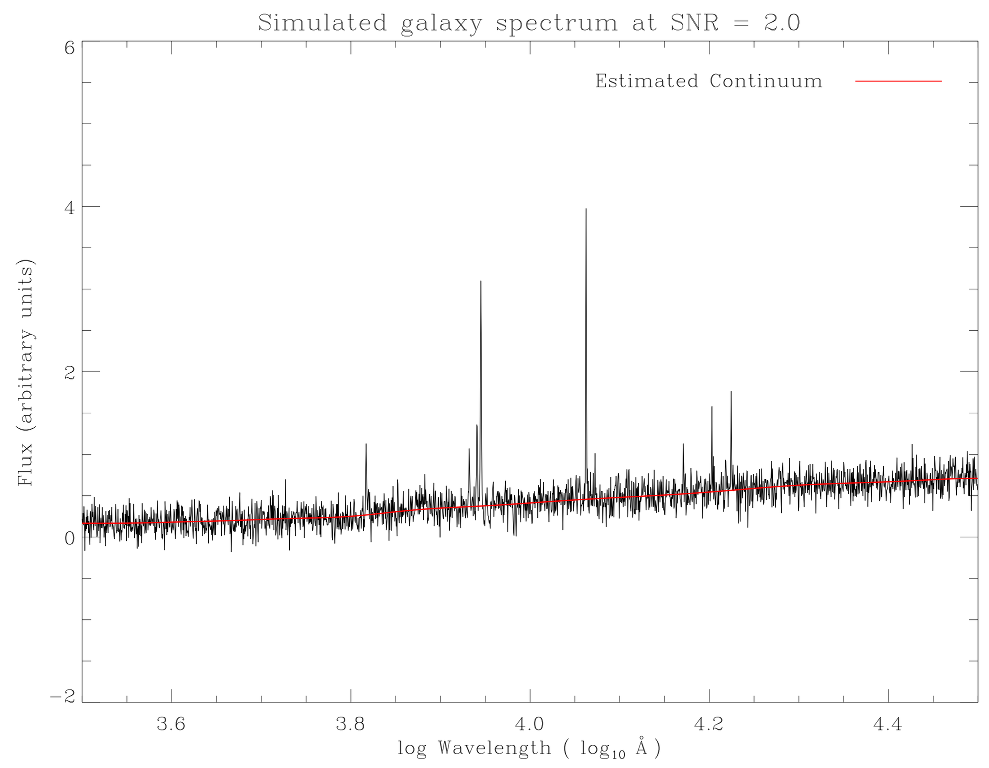
Blind Continuum Subtraction - An Example

- We begin with a pure noiseless spectrum, *S*, from our simulated catalogue, which consists of spectral lines, *L* and continuum, *C*, to which we add noise artificially.



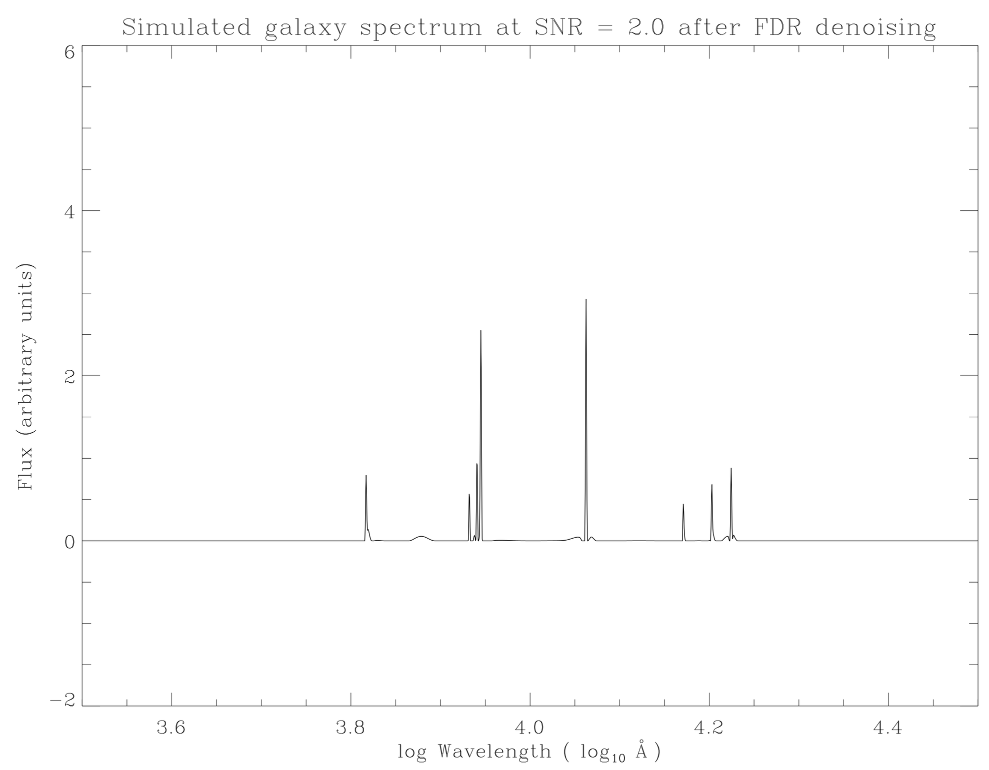
An example noiseless spectrum as obtained from our LePhare simulation.

- We add Gaussian noise, *N* on the spectrum. *S* can now be expressed as $L + C + N$.



The same galaxy spectrum with added Gaussian noise at a signal-to-noise of 2 on the spectrum. The estimated continuum is in red.

- After an application of mr1d_filter, including the removal of the last scale, we can obtain something that is approximately the lines only, L.

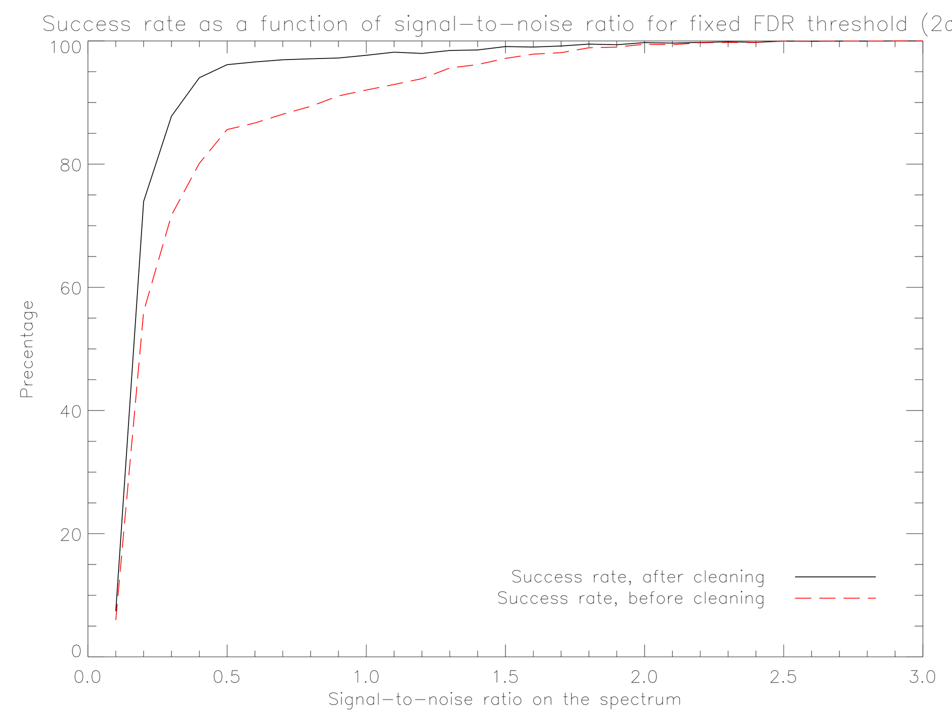


The FDR denoised spectrum, at 2σ thresholding, on the same scale. The principal lines are kept, but the very fine features are lost in the noise. This spectrum would be considered by Darth Fader as a good candidate for redshift estimation since it has > 3 peaks.

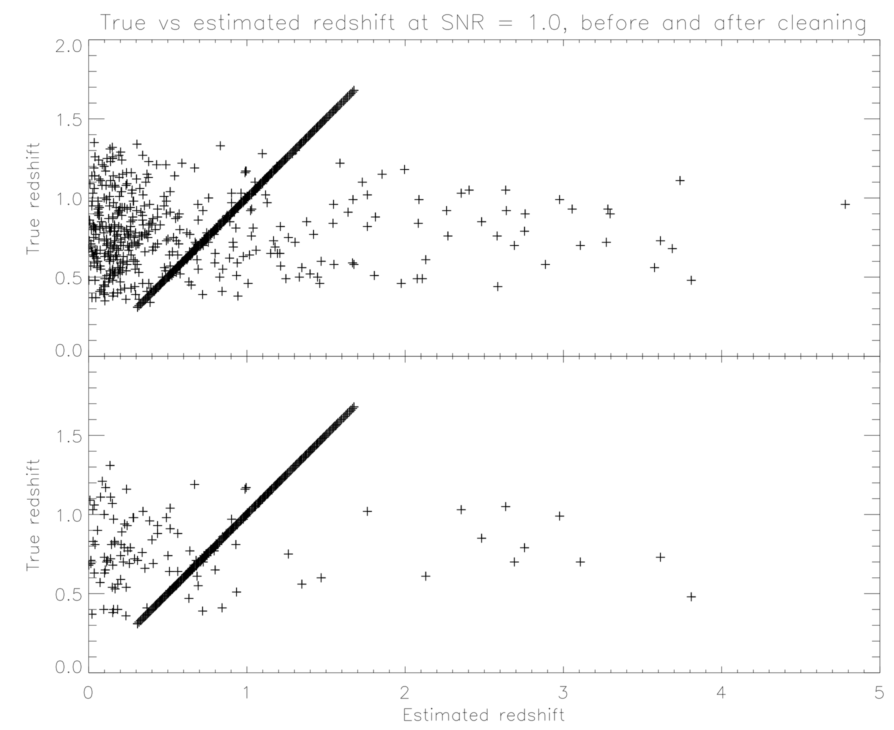
- We then subtract *L*, to get the continuum, plus the noise, $(C + N)$.
- Finally, we apply a wavelet filtering using star1d, and keep only the largest scale, which is analogous to the continuum, which we subtract off, keeping the *noisy* spectrum for the cross-correlation.

Results

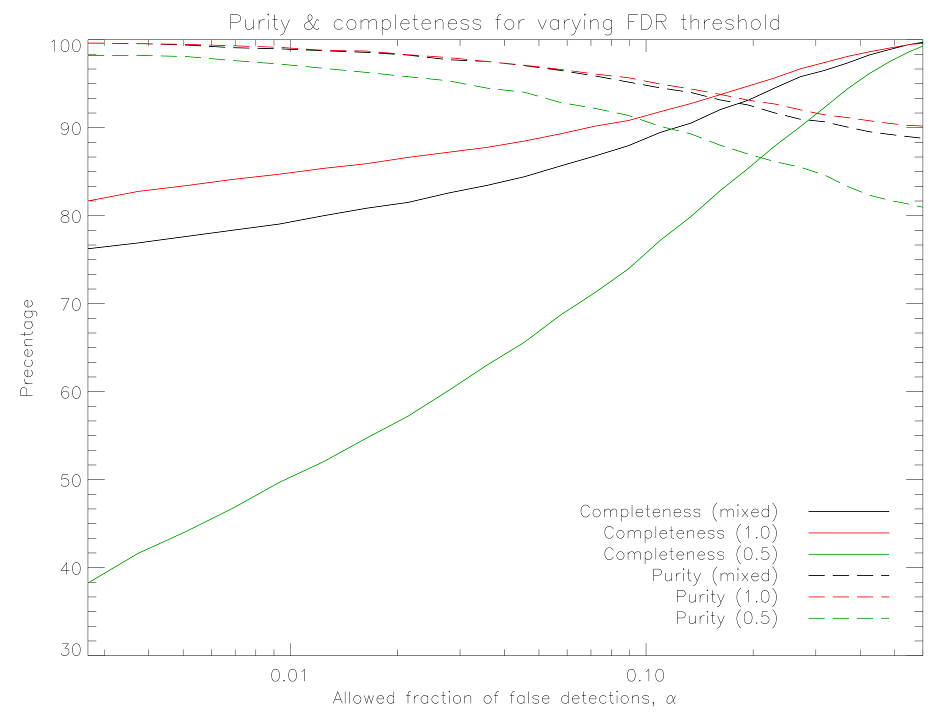
- By running Darth Fader on different catalogues we can compare the effect of varying signal-to-noise values & varying the FDR threshold. We compare the results with the application of cleaning via the 3-peak counting criterion, with those on the catalogue without cleaning.
- We define *success rate* as the percentage of redshifts that are correct to within $\Delta z = 0.01$; *completeness* as the percentage of galaxies retained after cleaning; and *purity* as the percentage correct relative to the total number of galaxies after cleaning.



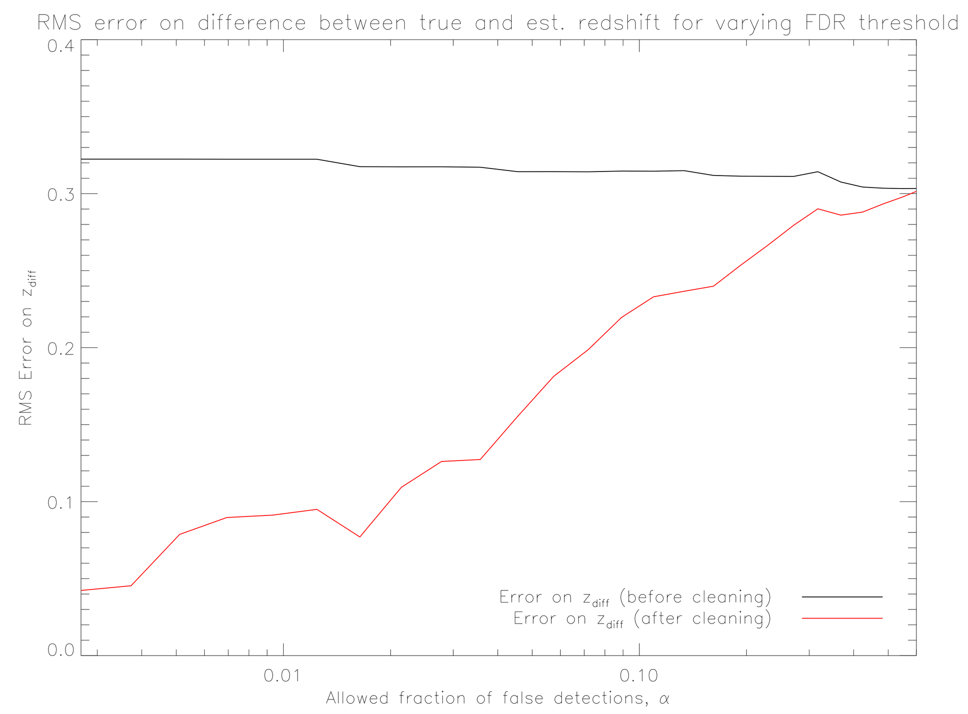
The impact of signal-to-noise level on the success rate achieved with an FDR denoising at a fixed value of 2σ (i.e.: $\alpha = 4.55\%$ allowed false detections). Note the marked improvement from Darth Fader in the SNR range 0.2 - 1.5.



A snapshot of how Darth Fader cleans a catalogue. Here, the signal-to-noise is at a value of 1, and the FDR threshold is again at 2σ. Outliers are significantly reduced, and the overall success rate is improved by $\sim$ 10%.



The performance of purity and completeness as a function of varying FDR threshold for two fixed SNR catalogues (at 0.5 and 1.0) and a further uniformly mixed SNR catalogue (range 0.1 - 3.0). Note the greater sacrifices required in completeness in order to obtain the same purity at an SNR of 0.5 compared to 1.0. Note also that we are able to obtain a 60% completeness rate for the galaxy catalogue at an SNR of 0.5 with a purity of over 95%



The effect of varying the FDR threshold on the RMS value of the difference between the true redshift from the simulation, and the estimated redshift from Darth Fader, for the mixed signal-to-noise catalogue. There is a clear decrease in RMS error with stricter thresholding. In principle, the upper line should be **constant** with a change in FDR threshold, however, since the same FDR thresholding is used in for the continuum subtraction procedures, it has a noticeable, but negligible, effect.

Discussion

- We have shown Darth Fader is a powerful tool for the improvement of redshift estimation without any *a priori* knowledge of galactic composition, type or morphology.
- Our approach is new in 2 key ways:
  1. We can *empirically* extract the continuum of a spectrum, without having to know any information about the physical processes that gave rise to it.
  2. We can make confident use of low signal-to-noise data, retaining high purity.
- We can segregate good data (data likely to yield accurate redshift estimates) from poor data empirically, with a high level of confidence.
- Even at signal-to-noise levels as low as 0.5 on the spectrum, attainment of 95% purity is possible whilst still retaining 60% of the data.
- Darth Fader offers the possibility of making practical use of a significant proportion of data that is currently thrown away. This represents a potential greater reach of spectroscopic surveys in terms of depth, since the faintest (and thus noisiest) galaxies in a survey will be primarily at higher redshift.
- The catalogues used consisted primarily of emission line simulated spectra, with artificial wavelength independent Gaussian noise. The catalogue would benefit from the inclusion of a more representative sample of galactic spectral types, and more realistic noise modelling.
- The mixed signal-to-noise catalogue is not representative of the distribution of SNR values of a real survey, which will be dependent on the nature of the survey, but is illustrative of the power of the Darth Fader method.
- Galactic spectra need to be wide enough in their wavelength range to encompass *at least* three main features.
- We expect that the Darth Fader algorithm can be improved by utilising photometric information in the two-peak cases – which are a significant source of completeness degradation.

References

- Arnouts, S., Cristiani, S., Moscardini, L., et al. 1999, Monthly Notices of the Royal Astronomical Society, 310, 540
- Coleman, G. D., Wu, C., & Weedman, D. W. 1980, Astrophysical Journal Supplement Series, 43, 393
- Glazebrook, K., Offer, A. R., & Deeley, K. 1998, Astrophysical Journal, 492, 98
- Ilbert, O., Arnouts, S., McCracken, H. J., et al. 2006, Astronomy & Astrophysics, 457, 841
- Kenobi, O.-W., et al. 1977, Alderaan Journal of Astrophysics, 20, 1038-1039
- Kinney, A. L., Calzetti, D., Bohlin, R. C., et al. 1996, Astrophysical Journal, 467, 38
- Machado, D. P., Leonard, A., Starck, J.-L., & Abdalla, F. 2012, *in prep.*
- Starck, J.-L. & Murtagh, F. 1994, Astronomy & Astrophysics, 288, 342

Acknowledgements

- The authors would like to acknowledge the support provided by the European Research Council, through grant SparseAstro (ERC-228261).

* http://jstarck.free.fr/isap.html

corresponding: daniel.machado@cea.fr