Darth Fader: A wavelet-based algorithm for empirical continuum subtraction of spectra, and flagging of poor quality spectra in cross-correlation redshift determination

D. P. Machado[1] *, A. Leonard[1], J.-L. Starck[1], F. B. Abdalla[2], and S. Jouvel[2,3]

[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.
[3] Institut de Ciències de l'Espai (IEEC-CSIC), E-08193 Bellaterra (Barcelona), Spain.

June 29, 2012

ABSTRACT

Context. Large sky surveys, and the sheer volume of data they produce, have made it necessary to tackle the problem of redshift identification in an automated and reliable fashion. Current methods attempt to do this with careful modelling of the spectral lines and the continua, or by employing a flux/magnitude or a signal-to-noise cut to the dataset in order to obtain reliable redshift estimates for the majority of galaxies in the sample.
Aims. In this paper, we present the DARTH FADER algorithm (**D**enoised and **A**utomatic **R**edshifts **Th**resholded with a **Fa**lse **D**etection **R**ate), which is a new wavelet-based method for estimating redshifts of galaxy spectra, and for the segregation of potentially incorrect redshift estimates from accurate ones. Automated, simple, and largely empirical, we explicitly highlight how the Darth Fader algorithm performs in the very low signal-to-noise (≤ 3) regime.
Methods. We employ a Principal Component Analysis (PCA) technique on a template set of zero redshift, high signal-to-noise simulated galaxies, and combine it with a new, model-independent method of continuum-subtraction involving a wavelet filtering procedure. This is followed by a standard cross-correlation and χ^2 minimisation procedure with a catalogue of simulated spectra (continuum subtracted in the same way) at nonzero redshifts. Finally a wavelet denoising with a false detection rate (FDR) threshold is applied to the redshifted spectra in order to obtain clean spectra. A peak counting criterion is then applied to these denoised spectra, where the presence of three or more peaks flags a spectrum as likely to yield an accurate redshift estimate. The noisy spectral catalogue is segregated on this criterion, with the spectra that fulfil it going on to be cross-correlated.
Results. We show that for this simple method and selection criterion, we can make use of an extremely low signal-to-noise catalogue, e.g.: with a signal-to-noise ratio of 0.5, with respect to the integrated flux of the spectrum, where it is still possible to recover 60% of the redshifts with a purity of 95%, allowing us access to data that would otherwise have been discarded. We also show that for a catalogue with uniformly mixed signal-to-noise ratios between 0.5 and 3.0, it is possible to recover 80% of the redshifts with 95% purity.
Conclusions. For very large sky surveys, this recovery of redshifts for low signal-to-noise spectra, even at less optimistic levels due to the added complexity of dealing with real data, would still represent a significant boost in the number of faint, and consequently high-redshift, galaxies with accurately determined redshifts.

Key words. Methods: data analysis – Techniques: spectroscopic – Galaxies: distances and redshift – Surveys

1. Introduction

The simplest method for estimating the redshift from a spectrum is by visual inspection, however, large sky surveys are providing astronomers with increasingly large datasets of spectra, necessitating the use of automated programs to obtain accurate information from the wealth of data available, as well as sophisticated techniques to remove the presence of noise, which becomes increasingly important for more distant, dimmer, sources.

Traditional methods for automated estimation of the redshifts of galaxy spectra have primarily been reliant upon template matching with cross-correlations (Tonry & Davis 1979; Glazebrook et al. 1998; Aihara et al. 2011) or – and sometimes in conjunction with – the matching of spectral lines (Kurtz & Mink 1998; Garilli et al. 2010; Stoughton et al. 2002).

Spectral line matching methods involve the use of spectra with a high enough signal-to-noise ratio to detect at least one emission line above a predefined threshold. With multiple emission lines, it is a task of matching the respective rest frame wavelengths of lines such as H_α and the O[III] doublet, to their respective redshifted counterparts. When faced with just a single emission line, assuming it to be H_α – since this is usually, but not always, the strongest feature – is a viable option for spectroscopic redshift determination of emission galaxies (Garilli et al. 2010). However, for additional constraints on what the lines can be, photometric data can be employed, as in the SDSS Early Data Release (Stoughton et al. 2002).

Template matching methods involve a catalogue of galaxy spectra at unknown redshift values being matched to a template set of spectra (intially at zero redshift) with some method of maximisation – usually a standard χ^2 test or a maximum likelihood estimate with additional priors – and accounting for the difference in redshift. Templates may come from simulated spectra based on modelling, local galaxy spectra whose redshifts are small and precisely known, or from high signal-to-noise spectra within the survey itself with redshifts that can be confidently identified.

Cross-correlation methods such as those described by Glazebrook et al. (1998) use a discrete Fourier transform

* e-mail: daniel.machado@cea.fr

and cross-correlation method to allow the shift of the template spectra (and thus redshift) to become a free parameter. Cross-correlation methods are useful because they can be computed as a multiplication in Fourier space between the template and galaxy spectra, resulting in easier and faster computation (via the Fast Fourier Transform or FFT) than performing the same procedure in real space.

Cross-correlations such as these, however, require the spectra to be free of continuum in order to correctly correlate line features with other line features and not continuum features. Currently, continua have to either be modelled from population synthesis models such as those of Bruzual & Charlot (2003), requiring *a priori* knowledge of galactic properties and physics, or they are computed through an averaging/subtraction process on galactic spectra of similar galaxy type (commonly via a PCA method), which again requires the *a priori* knowledge of how to identify and group galaxies which are of a similar type (Koski & Osterbrock 1976; Costero & Osterbrock 1977; Panuzzo et al. 2007). Polynomial fitting/statistical average methods are also frequently used when the noise is small enough so as not to conceal the continuum, or where denoising has already been employed, as in Stoughton et al. (2002); SubbaRao et al. (2002). In the very low SNR limit, it becomes exceedingly difficult to pinpoint exactly where the continuum lies, and polynomial fitting is not ideal.

In this paper we present a wavelet-based method that can isolate the continuum of a spectrum without having to defer to any knowledge of galaxy properties or physics, and that can operate in a high noise regime (§ 3). We also present a method for classifying which spectra are likely to give accurate redshift estimates. We do this by employing a denoising on the spectrum using a wavelet-filtering with a *false discovery rate*[1], henceforth FDR (Starck & Murtagh 2006), and counting the number of peaks in the clean spectrum. Evidently the more lines that are present, the greater the likelihood of finding a good match to a template. We choose the criterion of three peaks in the denoised spectrum to be a good match, by considering the resolution of the spectrum and how a matching may be done by eye – two lines presents an ambiguity that may not be resolvable, since, particularly on a log wavelength scale, some pairs of lines can be easily mistaken for other pairs with similar separation distances.

We generate mock catalogues (§ 2.2) using the simulation part of the LePhare program[2] (Arnouts et al. 1999; Ilbert et al. 2006) to generate noise-free template and galaxy catalogues. We create different noisy galaxy catalogues by adding random white gaussian noise to the original noiseless galaxy catalogue.

We implement a PCA decomposition of the template set (§ 2.1) to extract just 4 eigentemplates which then undergo cross-correlation with the galactic spectra in order to obtain redshift estimates.

We show results (§ 4) for varying signal-to-noise levels against a fixed FDR threshold, and the equivalent for fixed signal-to-noise and varying the FDR threshold. We also show the effects of varying the FDR threshold on a mixed signal-to-noise catalogue. We compare the success rates/purity and the completeness for our method with and without the FDR thresholding and peak-counting criterion that generates the galaxy subset that we retain.

We show that by appropriate choice of FDR threshold, it is possible to attain a high purity subset of the galaxy catalogue, whilst retaining a significant proportion of the population, even in the case of a signal-to-noise level of 0.5, where it is possible to obtain 95% purity with 60% completeness by choosing an FDR threshold of greater than 4.55% allowed false detections.

2. Method: Redshift Estimation from Spectra

2.1. Cross-Correlation and PCA

For Darth Fader we employ a PCA decomposition on a template set of galaxies following a similar derivation as in Glazebrook et al. (1998) with the difference of setting the weighting function, w_λ, and the normally distributed errors, σ^2_λ as wavelength independent and constant, and choosing to make the eigentemplates orthonormal.

Firstly we identify that any galaxy spectrum S_λ – initially at zero redshift, $S^{'}_\lambda$, where λ represents the wavelength bin – can be thought of as a vector in a high dimensional space, and we assume that it can be expressed as a linear combination of normalised template spectra, $T_{i\lambda}$, (also at zero redshift):

$$S^{'}_\lambda = \sum_i a_i \, T_{i\lambda} \, , \tag{2.1}$$

with normalisation:

$$\sum_\lambda T^2_\lambda = 1 \, . \tag{2.2}$$

If we choose to bin our spectra on a logarithmic wavelength axis, redshifting becomes proportional to a translation along the log-wavelength axis, which we label as Δ, where:

$$\begin{aligned} \Delta &= \log{(1+z)} \\ &= \log{\left(\lambda_{\,observed}\right)} - \log{\left(\lambda_{\,rest\ frame}\right)} \, . \end{aligned} \tag{2.3}$$

The estimate of the goodness-of-fit between the template, now allowed to shift along the wavelength axis, and the spectrum, at an unknown redshift, can be found by computing the minimum distance via a standard χ^2, where the previous coefficients, a_i are now dependent upon redshift, through Δ:

$$\chi^2(\Delta) = \sum_\lambda \frac{w^2_\lambda}{\sigma^2_\lambda} \left[S_\lambda - \sum_i a_i(\Delta) \, T_{i(\lambda+\Delta)} \right]^2 \, . \tag{2.4}$$

We can obtain values for each a_i per template, by maximising equation (2.4) with respect to $a_i(\Delta)$, where we set w^2_λ and σ^2_λ as wavelength independent and constant; giving:

$$a_i(\Delta) = \frac{\sum_\lambda S_\lambda \, T_{i(\lambda+\Delta)}}{\sum_\lambda T^2_{i(\lambda+\Delta)}} \, . \tag{2.5}$$

It should be noted that the numerator in equation (2.5) is a cross-correlation between the galaxy spectrum and

[1] We choose in this paper the synonymous, but more intuitive, term: false *detection* rate.

[2] Publicly available at: http://www.cfht.hawaii.edu/~arnouts/lephare.html

the i^{th} template spectrum. Substituting back into equation (2.4), we obtain:

$$\chi^2 \propto \sum_{\lambda} \left[S^{\,2}_{\,\lambda} - \sum_{i} a_i^2(\Delta)\, T^2_{i(\lambda+\Delta)} \right] . \qquad (2.6)$$

For a large and diverse galaxy catalogue, a substantial number of templates are needed ($\gtrsim$ 100) to ensure good coverage of all the complex galaxy types; to use all of them in the cross-correlation would render the method impractically long to compute.

Principal Component Analysis (PCA) is a simple tool that allows us to reduce the dimensionality of the problem. To do this we must extract the most important features from our templates – the principal components. The general procedure involves the construction and subsequent diagonalisation of a correlation matrix to find eigenvectors and eigenvalues. It is possible to construct a correlation matrix either between the templates, or between the wavelength bins, the result is equivalent. We have chosen to do the correlation between the templates since in our case the number of templates is less than the number of wavelength bins, resulting in a smaller matrix that is more amenable to inversion:

$$C_{ij} = \sum_{\lambda} T_{i\lambda}\, T^T_{j\lambda} . \qquad (2.7)$$

Since this correlation matrix is always real and square-symmetric, it follows that it is diagonalisable:

$$\mathbf{C} = \mathbf{R \Lambda R^T} , \qquad (2.8)$$

where $\mathbf{\Lambda}$ represents the matrix of ordered eigenvalues (largest to smallest) and $\mathbf{R}$, the matrix of correspondingly ordered eigenvectors. Finally, to get the *eigentemplates*, $\mathbf{E}$, we perform:

$$E_{\,j\lambda} = \frac{\sum_i R^T_{ij}\, T_{i\lambda}}{\sqrt{\Lambda_j}} , \qquad (2.9)$$

with the resulting templates having the same dimensions as the original dataset, and satisfying the orthonormality condition:

$$\sum_{\lambda} E_{\,i\lambda}\, E^{\,T}_{\,j\lambda} = \delta_{\,ij} . \qquad (2.10)$$

The effect of PCA is that it re-orientates the dataset to lie along the orthogonal eigenvectors (axes) sorted by descending variance. It is assumed that the eigenvector with the greatest variance (largest eigenvalue) corresponds to strongest signal features of the untransformed dataset, with subsequent eigenvectors representing less significant signal features, and the final eigenvectors, with the smallest variances, representing noise. Since PCA creates an 'importance order', it is often possible to associate the first few principal components with physical features; for example, if H_α is a very prominent feature in most of the template spectra, it will be present in at least the first principal eigentemplate. With this in mind we can now re-cast equation (2.1) in terms of an approximation of the sum of the first N eigentemplates that are now allowed to be shifted along the wavelength axis:

$$S_{\,\lambda} \simeq \sum_{i=1}^{N} b_i(\Delta)\, E_{i(\lambda+\Delta)} , \qquad (2.11)$$

where $b_i(\Delta)$ are new expansion coefficients for the new basis.

Equations (2.5) and (2.6), using the orthogonality condition from equation (2.10) then become:

$$b(\Delta) = \sum_{\lambda} S_\lambda\, E_{(\lambda+\Delta)} , \qquad (2.12)$$

$$\chi^2 \propto \sum_{\lambda} S^{\,2}_{\,\lambda} - \sum_{i=1}^{N} b_i^2(\Delta) . \qquad (2.13)$$

We then observe that the first term in equation (2.13) is not affected by redshifting the eigentemplates and as such is a constant in the χ^2 function, and can be disregarded; therefore we now wish to *maximise* the related function, $\widetilde{\chi}^{\,2}$, since the χ^2 function will then be at a corresponding minimum:

$$\chi^2 \sim \widetilde{\chi}^{\,2} = \sum_{i=1}^{N} b_i^2(\Delta) . \qquad (2.14)$$

Hence, for each eigentemplate E_i, finding the $b_i(\Delta)$ term is a case of cross-correlating the galaxy spectrum with the i^{th} eigentemplate spectrum. And the total value of the $\widetilde{\chi}^{\,2}$ equation for any particular galaxy will be the sum of the cross-correlations with each of the N eigentemplates that have been retained.

We can further simplify the problem by noting that a convolution between two real signals transforms into a multiplication in Fourier space between the individual Fourier transforms of the galaxy and non-redshifted template spectra, with the desirable aspect of Δ becoming separate from the eigentemplates, and thus a free parameter. Hence we obtain:

$$b_i(\Delta) = \mathcal{F}^{-1}\big(\hat{S}_k\, \hat{E}_{ik}\big) = \frac{1}{M} \sum_{k=0}^{M-1} \hat{S}_k\, \hat{E}_{ik}\, e^{\frac{2\pi \mathrm{i} k \Delta}{M}} , \qquad (2.15)$$

and:

$$\widetilde{\chi}^{\,2}(\Delta) = \sum_{i=1}^{N} \Big[\mathcal{F}^{-1}\big(\hat{S}_k\, \hat{E}_{ik}\big) \Big]^2 , \qquad (2.16)$$

where $\hat{S}_k$, $\hat{E}_{ik}$, represent the Discrete Fourier Transforms (DFTs) of S_λ, $E_{i\lambda}$; and i, $\mathcal{F}^{-1}$ represent $\sqrt{-1}$ and the inverse DFT respectively.

Now that we have obtained equation (2.16) it is an easy task to extract the estimate for the redshift, z. 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:

$$z_{est} = 10^{\,\delta_s\, \Delta_{\widetilde{\chi}}} - 1 , \qquad (2.17)$$

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

Note that, for this PCA/cross-correlation redshift estimation method, both the template and galaxy spectra must be free of continuum. This is important to ensure that it is the spectral features that are being matched, rather than the continuum, which may lead to spurious correlations and hence confusion in the determination of the galaxy redshift. In Darth Fader, we use an entirely empirical method for subtracting the continuum that is based on the wavelet decomposition of the spectrum, and which is easily automated and very effective regardless of the signal-to-noise of the spectrum under consideration. This method will be described in detail in § 3.3.

2.2. Mock Catalogue

In this paper, we use a catalogue of simulated galaxy spectra to test and verify the effectiveness of our method. In addition we use simulated galaxies at very high signal-to-noise to generate our template set, which is then used to determine the principle components for cross-correlation with the simulated galaxy spectra in order to estimate their redshift, as described above.

To generate these mock catalogues we used the simulation part of the LePhare program (Arnouts et al. 1999; Ilbert et al. 2006) with 3,000 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. The filters evenly span the $\log_{10}$ of the wavelength axis from 3.0 to 4.5, corresponding to a wavelength range of 1,000 Å to 31,623 Å. The galaxy catalogue was made in a similar manner, but using only 2,000 top-hat filters, resulting in a shorter wavelength range of 3,160 Å to 31,623 Å (corresponding to a range of 3.5 to 4.5 on the log axis). This choice of binning gives a constant resolution across the spectrum of $R\left(={}^{\lambda}/_{\Delta\lambda}\right) \sim 870$ for all the catalogues, and a grid/pixel spacing of $\delta_s = 5 \times 10^{-4} \log_{10}$ Å.

The template set consisted of 259 galaxies, and the galaxy catalogue of 3,775, with the CWW-Kinney standard templates (Coleman et al. 1980; Kinney et al. 1996) being used throughout.

Different SNR catalogues (from SNR = 0.1 through to SNR = 5.0) of the galaxy spectra were created manually by adding wavelength independent (white) gaussian noise to the spectra. The signal-to-noise level is defined relative to the average integrated flux over all the spectra in the noiseless galaxy catalogue. An additional catalogue was created with mixed SNR values, uniformly spread from 0.1 to 3.0, again with gaussian noise.

3. Sparse Wavelet Spectral Analysis

Current methods for continuum-subtraction rely on one of two principal techniques: careful modelling of the physics of galaxies to estimate the continuum in order to subtract it off; or an averaging of a set of similar spectra (same galaxy type) which is then subtracted off or divided out, (Koski & Osterbrock 1976; Costero & Osterbrock 1977; Panuzzo et al. 2007).

Both these methods have the undesirable quality of requiring some knowledge of galaxy physics, and being somewhat restricted to lower redshift galaxies. Careful modelling is computationally intensive and liable to failure if unusual galaxy types are found. Averaging methods require a priori knowledge of the galaxy type of a set of galaxies. Both methods generally do not inherently account for evolutionary differences of galaxies at greater redshifts, though this can be added in. A further alternative that is limited to high signal-to-noise spectra, or spectra that have been denoised beforehand is polynomial fitting (as applied to the SDSS data release, Stoughton et al. (2002)). By contrast our new method of subtracting off the continuum is completely *empirical*, and requires no knowledge of the physics or type of the galaxy involved and can be used even with very noisy spectra. We do this by using a combination of a wavelet transform and a wavelet denoising.

3.1. The Discrete Wavelet Transform

A discrete wavelet transform (DWT) is a method of decomposing a signal (in this case, a finite and real 1D signal) onto a new basis that is defined by a mathematical object known as a wavelet. This transformation is simply a new representation of the original signal; the signal itself is not altered in any way, and can be recovered through a simple summation.

The DWT is somewhat similar to a discrete Fourier transform (DFT), where one can express any function/signal as being decomposed on a finite basis (alternatively a finite sum) of sines and cosines, with varying amplitudes (expansion coefficients) and frequencies. A DWT is similar in the sense that there exists a well defined (but not unique) mathematical object - the wavelet - which plays an analogous role to the sine/cosine basis in a Fourier transform. However, this expansion on a wavelet basis replaces the role of frequencies with a finite number of scales that pick out varying features within the signal (which is something that a Fourier transform is not optimal for) ranging from slowly changing, to rapidly changing, with the finesse of the variation being defined by the number of scales chosen. The largest scale corresponds to the slowest varying component of the signal.

A useful analogy to help convey the role of the wavelet scale, is to think of the scale on a map of a city. A map with a very large scale gives the broad overall picture of the terrain and where the main regions of the city are; conversely a map with a very small scale allows you to see all of the finer detail of the small streets and buildings.

In this paper, we use the isotropic undecimated wavelet transform (IUWT), also called the *starlet wavelet transform*, which is well known in the astronomical domain because it is well adapted to astronomical data where, to a good approximation objects are commonly isotropic (Starck & Murtagh 1994, 2006).

Using the starlet transform, a spectrum of n bins, $S_\lambda = S\left[1,\ldots,n\right]$ can be decomposed into a coefficient set, $W = \{w_1,\ldots,w_J,c_J\}$, as a superposition of the form:

$$S_\lambda = c_J(\lambda) + \sum_{j=1}^{J} w_j(\lambda)\ , \tag{3.1}$$

where c_J is a smoothed/coarsely resolved version of the original spectrum S_λ, and the w_j coefficients represent the details of S_λ at scale 2^{-j}; thus, the algorithm outputs $J+1$ sub-band arrays each of size n.[3]

3.2. Treatment of Wavelet Coefficients

A useful application that has been derived from the properties of wavelets is that of denoising noisy data. Using the properties as described above, it is possible to clean a signal of considerable noise, whilst still retaining the vast majority of the information contained in the signal. This procedure is not perfect however, and occasionally – particularly with higher noise, or when some signal features are comparable to the noise level – some information in the signal can be lost along with the noise removal.

This denoising is achieved by applying a wavelet transform to the signal and then imposing some form of constraint on the wavelet coefficients – for example, considering only the coefficients above a certain (hard) threshold – and then reconstructing the signal. This procedure would remove the smallest coefficients below the set threshold, which if chosen appropriately would correspond mostly to noise, and hence the noise would be filtered from the signal. This is a basic example of wavelet denoising.

Many different methods exist for the treatment of the resulting wavelet scales once a signal has undergone a DWT. They range from simple hard thresholding and discarding the smallest wavelet coefficients as in a $K\sigma$ clipping, to more complex and sophisticated methods such as denoising with FDR which makes it possible to control contamination from false positive lines arising from noise features, in the subsequent denoised signal.

Wavelet denoising has been previously applied successfully to both stellar (Fligge & Solanki 1997; Lutz et al. 2008), and galactic spectra (Stoughton et al. 2002, for the SDSS early data release).

3.3. Sparse Wavelet Modelling of Spectra

We now consider how a spectrum can be decomposed into three components (continuum, emission/absorption lines, and noise), by using a modified version of a specific denoising algorithm based on the hybrid steepest descent (HSD) minimisation algorithm developed by Yamada (2001).

As an example, we show in figure 1a a typical noise-free spectrum from our simulated catalogue. In figures 1b and 1c, we show the same spectrum with noise added such that the SNR of the spectrum is 2 and 0.5, respectively. Overplotted in these latter two figures is the continuum as estimated by the method described below.

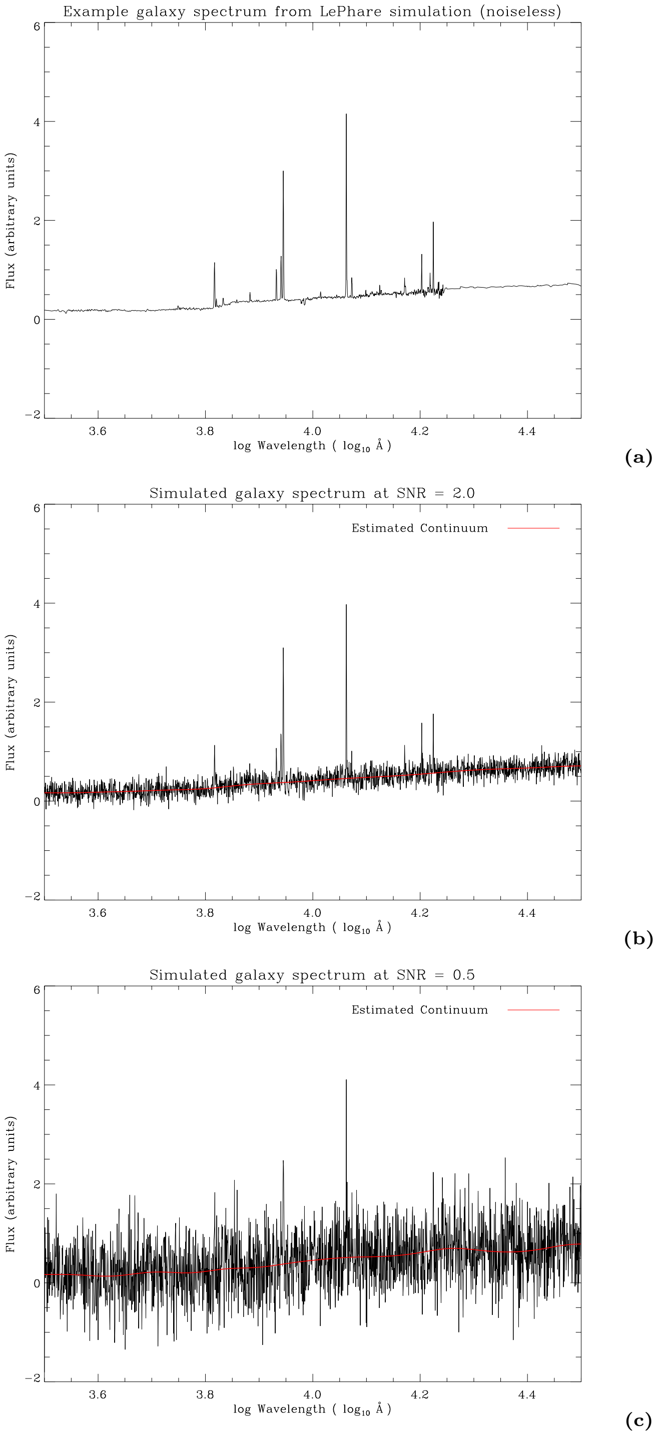


Fig. 1: Figure 1a shows a typical spectrum from the galaxy catalogue, as generated by a LePhare simulation, without any noise. Figures 1b and 1c are the same spectrum as that in figure 1a but with manually added noise at a signal-to-noise level of 2.0 and 0.5 on the spectrum respectively. The red lines indicate the empirically determined continuum in both cases. Note that the majority of the prominent lines are easily visible by eye at an SNR of 2.0, whereas at an SNR of 0.5 only one line (the most prominent one in the original) is unambiguously identifiable by eye, with the second most prominent line also being visible, but of a similar size to a noise feature located at approximately 4.37 on the log wavelength axis.

It can be seen from the figures that a measured spectrum is comprised of three separate components; we therefore can model the observed spectrum S, as the following linear sum:

$$S = L + N + C\,, \tag{3.2}$$

where L is the unknown noise-free and continuum-free spectrum (i.e. only emission/absorption lines) , N is the noise and C is the continuum.

Assuming that L is composed of set of emission/absorption bands, not necessarily of the same width,

[3] IDL routines *stard1d.pro* and *istar1d.pro* for wavelet transformation and reconstruction are included in the **iSAP** package available at: `http://jstarck.free.fr/isap.html`. For a detailed description of the starlet transform see Starck et al. (2010). The present indexing is such that $j = 1$ corresponds to the finest scale or highest frequencies.

and that the continuum does not contain small scale features, we can consider that non-coarse wavelet coefficients $W^{'} = \{w_1, \dots, w_J\}$ can properly represent the emission/absorption bands, while information relative to the continuum is contained in the coarse resolution coefficient, c_J. Hence, we can first reconstruct L solving the following constrained convex optimisation problem:

$$\min_{L} \left\| \hat{\mathcal{W}} L \right\|_1 , \qquad \text{s.t.} \qquad S \in \mathcal{C}, \tag{3.3}$$

where $\hat{\mathcal{W}}$ is the wavelet transform operator, and $\mathcal{C}$ is a closed convex set of constraints including the linear data-fidelity constraints:

$$\left| w_j^{[S]}(\lambda) - w_j^{[L]}(\lambda) \right| \le \varepsilon_j, \; \forall \, (j, \lambda) \in \mathcal{M} \; , \tag{3.4}$$

where $w_j^{[S]}$ and $w_j^{[L]}$ are respectively the wavelet of coefficients of S and L, and ε_j is an arbitrarily small parameter. Note that the coarse scale coefficients of S are not considered in this minimisation. Therefore, we don't expect the continuum information to be included in the solution L. $\mathcal{M}$ is the *multiresolution support* (Starck et al. 1995), which is determined by the set of detected significant coefficients at each scale j, and location (wavelength bin) λ, as:

$$\mathcal{M} := \{(j, \lambda) \mid \text{if } w_j(\lambda) \text{ is declared significant}\} \; . \tag{3.5}$$

The multiresolution support is obtained from the noisy data S by computing the forward transform coefficients $W = \{w_1, \dots, w_J, c_J\}$, and recording the coordinates of the coefficients w_j with an absolute value larger than a detection level threshold τ_j, generally chosen as $\tau_j = K\sigma_j$, where K is user parameter (typically between 3 and 5) and σ_j is the noise standard deviation at scale j. An interesting and more efficient alternative to this standard $K\sigma$ detection approach is the procedure to control the False Discovery Rate (FDR), i.e.: the average fraction of false detections over the total number of detections (see Starck & Murtagh 2006, for more details).

The minimisation can be achieved using a version of the HSD algorithm adapted to non-smooth functionals. HSD allows minimising smooth convex functionals over the intersection of fixed point sets of non-expansive mappings. More details can be found in Starck et al. (2010).

Figures 2 and 3 show the reconstructions of the lines, L, from figures 1b and 1c, using an FDR threshold of 4.55% allowed false detections.

Once L is estimated, we can compute the residual $\Omega = S - L$, which contains only the continuum and noise. The largest scale of the residual contains the continuum information. One way to estimate the continuum consists in first taking the wavelet transform of Ω, i.e.: $W = \hat{\mathcal{W}}\Omega = \{w_1, \dots, w_J, c_J\}$, and then retaining only the largest scale: $c_J = C$.

3.4. The Darth Fader Decomposition Algorithm

The final algorithm to decompose a spectrum into its three main components[4] (i.e. continuum, emission/absorption lines, and noise) is the following:

1. Compute the wavelet transform $w_j^{[S]}(\lambda)$ of the input spectrum S_λ.
2. At each scale j, derive the threshold τ_j using the FDR approach.[5]
3. Compute the multiresolution support $\mathcal{M}$.
4. Solve equation (3.3) using the HSD algorithm to estimate L.
5. Compute the residual, $\Omega = S - L$.
6. Compute the wavelet transform of the residual, $W = \hat{\mathcal{W}}\Omega = \{w_1, \dots, w_J, c_J\}$ of Ω.
7. Estimate the continuum C with $C = c_J$.
8. Estimate noise N with $N = S - L - C$.
9. Compute the continuum-subtracted spectrum $S_{cs} = S - C$.

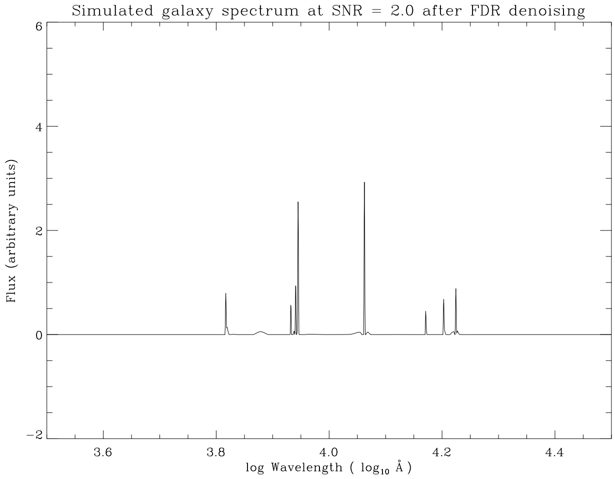


Fig. 2: This figure is the result of denoising the spectrum in figure 1b with an FDR threshold corresponding to an allowed FDR of 4.55%. Essentially all of the principal features from the pure spectrum (figure 1a) are clearly visible. No false detections are present.

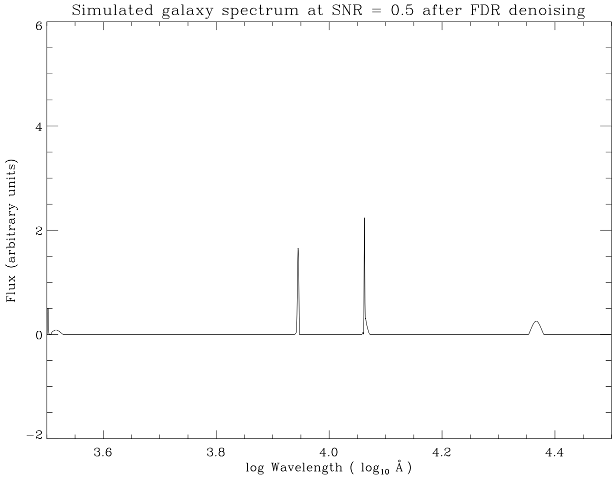


Fig. 3: The same denoising as applied in figure 2, now applied to figure 1c. Note the significant deterioration of detected (true) lines compared to when the signal-to-noise level was higher (fig. 2), and in particular, note the appearance of spurious detections at the beginning of the spectrum and at the previous line-like noise feature identified in figure 1c. The peak-counting algorithm we use would count this spectrum to have 5 peaks, and as such it would be kept as potentially good data for cross-correlation.

[4] Darth Fader is publicly available at: http://jstarck.free.fr/isap.html.

[5] It is possible to use $K\sigma$ thresholding, however, this would not be as robust to the presence of high noise.

The end result of the Darth Fader decomposition algorithm for the previous example spectra are shown in figures 4 and 5.

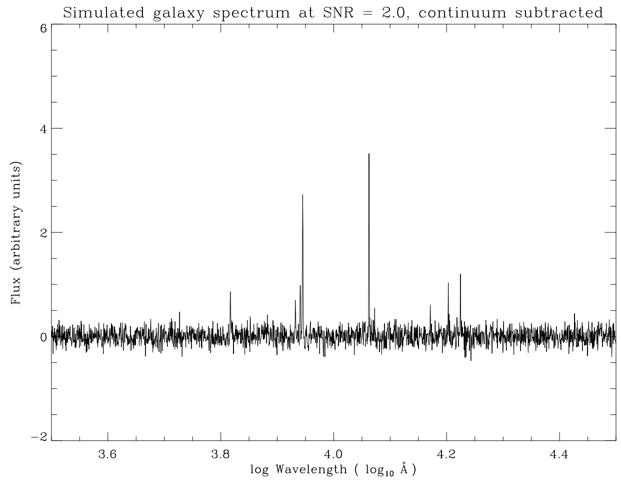


Fig. 4: The spectrum at a signal-to-noise level of 2.0, with the empirically determined continuum (as shown in fig. 1b) subtracted.

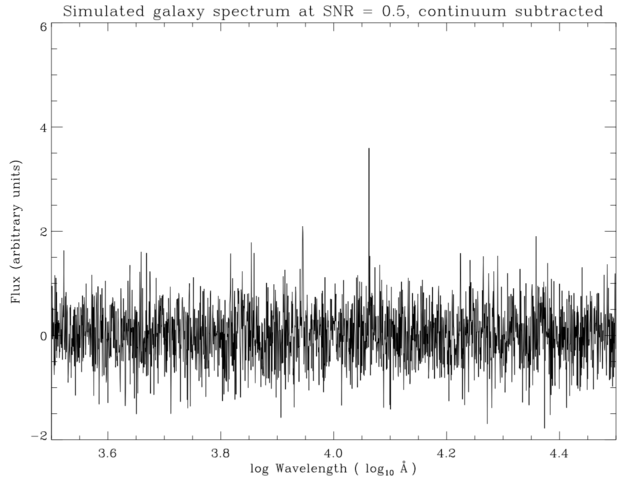


Fig. 5: This figure shows the spectrum in fig. 1c, at a signal-to-noise of 0.5, with the continuum subtracted.

3.5. Redshift Estimation

For our method, we follow the PCA procedure as described in § 2.1 on a set of very high signal-to-noise template spectra (SNR > 1,000) to obtain eigentemplates, of which we keep only the first four principal components. We continuum-subtract our spectra as described in section § 3.3. Since the noise on a spectrum will in principle be uncorrelated with the eigentemplates, we choose to use the noisy galaxy spectra in the cross-correlation – thus preserving any line information that may be hidden in the noise.

Evidently, at low SNR, some of these cross-correlations will produce erroneous results, and hence, a method of classifying when an estimate is likely to be a correct one is needed.

Consider if the redshift of a noisy galaxy spectrum were to be estimated by eye. Initially, a search for prominent lines above the background noise would take place. If one line is found, it is difficult to know which line it is, in particular if the wavelength range on the observed spectrum is narrow. If two lines are found, it now becomes easier to guess, though there is still significant potential for error since some distances between lines on a log-wavelength axis are of a very similar size – notably, O[II], O[III], H_α, and S[III], (rest frame wavelengths ~3,727, 5,007, 6,563, 9,069 Å respectively; see figure 1a) – and if not all of them are present, it becomes difficult to know which pair is presented, since there is no relative flux information to serve as a method for distinguishing the lines. If three lines are found, it should now be possible to distinguish with more reliability which emission lines are present in the spectrum, and thus, what the likely redshift is.

With this is mind, we choose a very simple, yet effective, criterion to decide whether the redshift estimate of a galaxy spectrum is likely correct. We employ a further FDR denoising on the continuum-subtracted spectrum (§ 3.3) and identify the number of peaks present via a simple peak-counting algorithm[6]. We then segregate the catalogue such that if a denoised spectrum presents 3 or more peaks, we keep the redshift determination as likely to be accurate, whereas should the spectrum possess only 2 peaks or fewer, we discard it and assume the redshift estimate to be unreliable.

Experimental tests showed that setting the selection criterion to two lines or more is indeed an insufficient discriminator, with the redshift estimates being contaminated by line confusion, whereas four lines proves to be too strict a criterion – much data is lost in this case, with no noticeable improvement in the percentage of retained galaxies with the correct estimated redshift. This criterion stems in part from the resolution of the spectra – if the resolution is poor, peaks are more coarsely resolved and thus occupy a certain width about their locations, subsequently making it easier to confuse lines since their precise locations in wavelength are slightly smeared out. The higher the resolution, the more narrowly resolved the lines become, and hence the more well located in wavelength. It should be evident therefore, that for spectra with higher resolutions, it becomes possible to relax the number of peaks required to two; for poorly resolved spectra, an increase in detected peaks would be needed.

Peaks are considered to be anywhere where the derivative of the spectrum changes from positive to negative (maxima), but only in the spectrum's positive domain; this means that, for example, a gaussian-like function with two maxima (a line-doublet), would count as *two* peaks. This method ignores absorption features, however, it can be easily adapted to also deal with them by counting the peaks on the absolute value of the FDR denoised spectrum.

Clearly, the FDR denoising can pick out more features than the human eye can see, therefore in an example where only one peak may be visible in the noisy spectrum, the FDR denoised spectrum may pick out 3 peaks. And hence we can obtain information on the reliability of the redshift when previously it was not known. Occasionally the FDR denoising will detect lines that are not true spectral fea-

[6] Algorithm adapted from 'peaks.pro', available from: http://astro.berkeley.edu/~johnjohn/idlprocs/peaks.pro

tures, however, a maximum for this false line contamination is set by the FDR threshold, α.

At low SNR there is a trade-off between relaxing the threshold to pick up more features – or indeed any features[7] – and imposing a stricter threshold to prevent the detection of spurious lines. Therefore if a spectrum presents 3 peaks, it is possible that some of them will be spurious, and this will lead to an erroneous redshift estimate from cross-correlation, and false-positive contamination of our retained data. Also, of the spectra that present with only two lines, some will have lines that do not suffer from confusion with any other pair of lines, and hence the redshift estimate would otherwise be reliable; the criterion chosen leads them to be discarded giving rise to false-negative contamination of the discarded data.

4. Experimental Results

For the purposes of illustrating the method, we ran Darth Fader over multiple FDR thresholds whilst keeping the signal-to-noise constant; and again over multiple signal-to-noises whilst keeping the FDR threshold constant; and finally Darth Fader was run with different FDR thresholds on a uniformly mixed SNR catalogue with SNR values ranging from 0.1 to 3.

We define the completeness, success rate, purity and loss of the sample to be, respectively:

$$C = \frac{R}{T} \times 100\% \,, \tag{4.1}$$

$$S = \frac{(T-F)}{T} \times 100\% \,, \tag{4.2}$$

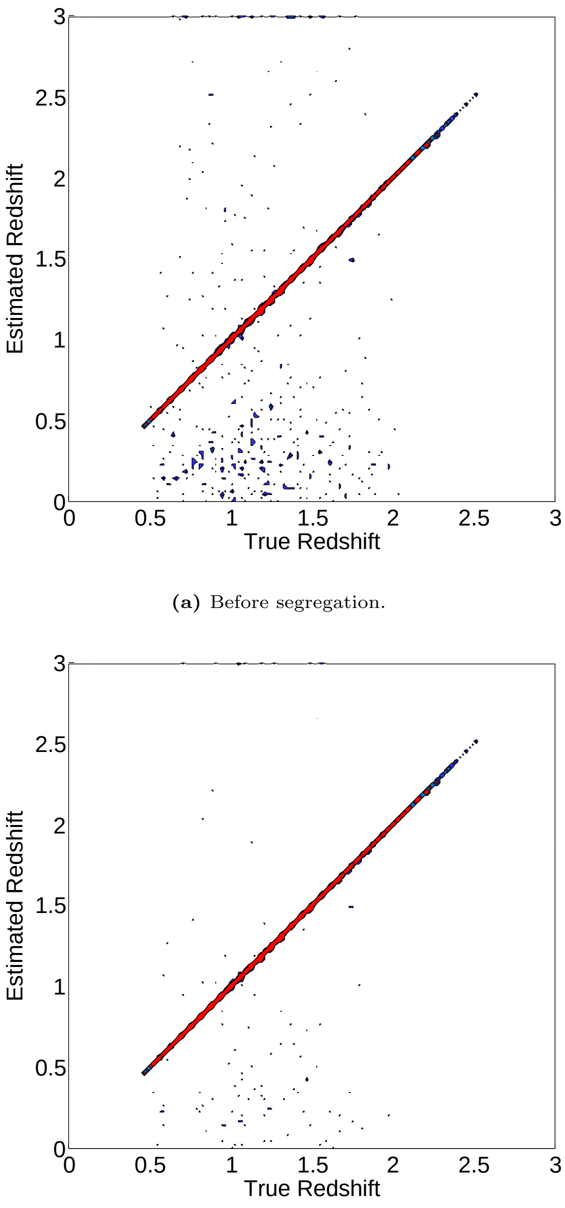
$$P = \frac{(R-F^{+})}{R} \times 100\% \,, \tag{4.3}$$

$$L = \frac{F^{-}}{T} \times 100\% \,, \tag{4.4}$$

where T and R respectively denote the total number of galaxies in the sample (before segregating) and the retained number of galaxies (the number that satisfy the 3-peak criterion). F denotes the number of failures in T, and F^{+} the number of false positive contaminants in R, (the number of galaxies that are kept by the Darth Fader algorithm, but still produce an incorrect redshift estimate); F^{-} denotes the number of false negatives, i.e.: the spectra that would have given the correct redshift when cross-correlated, but failed to satisfy the three-peak criterion and were thus discarded. The success rate equation (4.2), as shown in the example in fig. 7, is thus an analogous statistic to purity (equation (4.3)) except that it is on the global set without any catalogue segregation having taken place.

We present the result of segregating the catalogue of potentially poor redshift estimates, at an FDR threshold of 4.55% allowed false detections on one catalogue at an SNR of 1.0 in figure 6. This contour plot demonstrates a single realisation of how Darth Fader works.

Figure 7 shows the success rate of Darth Fader before and after catalogue segregation for a fixed FDR threshold corresponding to 4.55% allowed false detections when cleaning the spectra, as a function of varying signal-to-noise on the spectrum. As expected at extremely low levels of signal-to-noise both methods converge (to zero) since there is essentially no extractable information left in the spectrum as it has been completely dominated by noise. At high enough signal-to-noise levels, (in this case at about 2.3) both methods again converge, at 100% success rate, since denoising does not reveal any more useful diagnostic information – the number of clear peaks present in the noisy spectrum



(a) Before segregation.

(b) After segregation.

Fig. 6: A density plot to show the effect on redshift estimation before and after the segregation of poor data. Progression from red to blue represents a progression from high to low density. Figure 6a depicts the results before segregation, and 6b, after. A small number of erroneous, high estimated redshifts (i.e.: $z > 3$) are cumulatively binned at the upper border of the 6a. Clearly outliers still exist after segregation of the catalogue, where the redshift estimation has failed, but as it can be seen, a significant proportion of misclassified data is eliminated.

[7] If the noise level on the spectrum is sufficient to completely dominate over the signal or the FDR threshold particularly strict, the FDR denoising will fail altogether.

is *already sufficient* to obtain accurate redshift estimates from cross-correlation.

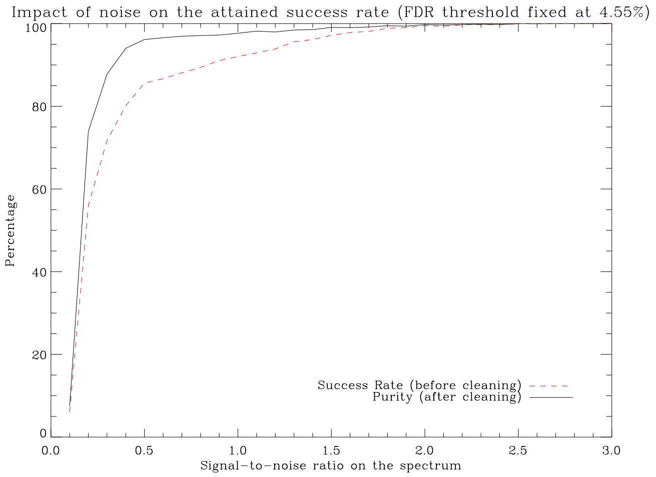


Fig. 7: A figure to show how Darth Fader improves the success rate of the redshift estimates of the galaxy catalogue at different signal-to-noise values for a fixed FDR threshold of 4.55% allowed false detections. Note the marked improvement in the SNR range 0.2 - 1.5.

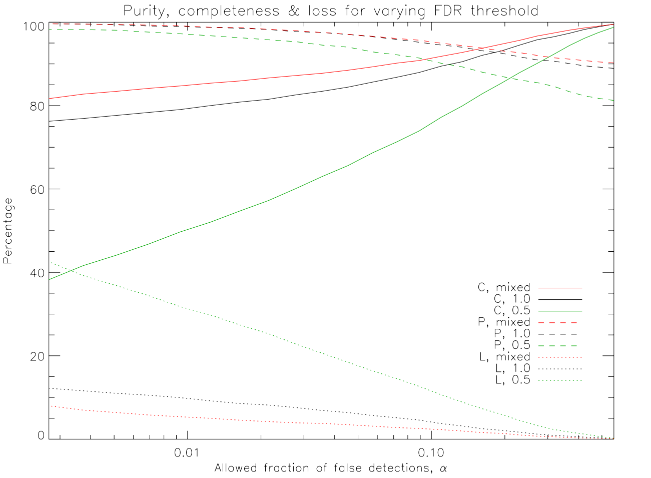


Fig. 8: A figure to show how varying the FDR threshold affects purity, completeness and loss at fixed signal-to-noise catalogues of 0.5 and 1.0, and a further uniformly mixed SNR catalogue. Note the greater sacrifices required in completeness in order to obtain the same purity at an SNR of 0.5 compared 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%.

To demonstrate the effects of the FDR threshold on purity and completeness we test Darth Fader on two fixed SNR catalogues of SNRs 0.5 and 1.0; and one mixed SNR catalogue, with uniformly spread SNR values from 0.1 to 3.0 (figure 8). Notice that beyond a certain point the purity saturates, and stricter FDR thresholding (resulting in a smaller fraction of false detections) does not result in a significant gain in purity, and indeed, only serves to penalise completeness.

Figure 8 clearly demonstrates the tradeoff that exists between a relaxation of the threshold (i.e.: increasing α) or a stricter application (decreasing α). Relaxing the threshold improves both the completeness and loss rate by detecting more features in the spectra – though not necessarily true features – thereby increasing the number of spectra accepted under the 3-peak criterion, but as a cost this penalises the purity.

A more conservative approach leads the FDR denoising to progressively remove real features, with the guarantee that very few of the remaining features will be false detections. This leads to a decrease in the completeness and an increase in the loss rate, since fewer spectra will exhibit 3 features after denoising, irrespective of whether they will subsequently produce correct redshift estimates, with the benefit of this being a boost in purity, and the knowledge that what is kept will very likely have the correct redshift.

The results of the uniformly mixed SNR catalogue represents a step toward a more realistic view of what a galaxy survey could look like. The restricted range of low SNR values in the mixed catalogue, by performing much closer in terms of purity and completeness to a fixed catalogue of SNR = 1.0, demonstrates that for a catalogue with a more extended range to higher SNR, this result would be a minimum. However, a real survey would not have this uniform distribution of SNR values, and would be skewed toward a greater population at lower SNR, therefore stifling the gain from an increased SNR range. The actual distribution in signal-to-noise of the galaxy catalogue will be specific to the type of survey and the instrument used.

Finally, we compare the RMS error on $z_{diff} = |z_{true} - z_{estimated}|$, in figure 9.

In principle, the upper line (before segregation) should be invariant with a change in FDR threshold, however, since the same FDR thresholding is used in both the peak-counting *and* the continuum-subtraction procedures, it has a noticeable, but negligible, effect. Consequently, the profile of the lower line (after segregation) is a composite one, and accounting for the continuum contribution, we should expect that the RMS error purely down to catalogue segregation would be marginally reduced. Although there is some variability due to the random nature of the noise, stricter thresholding improves the RMS error.

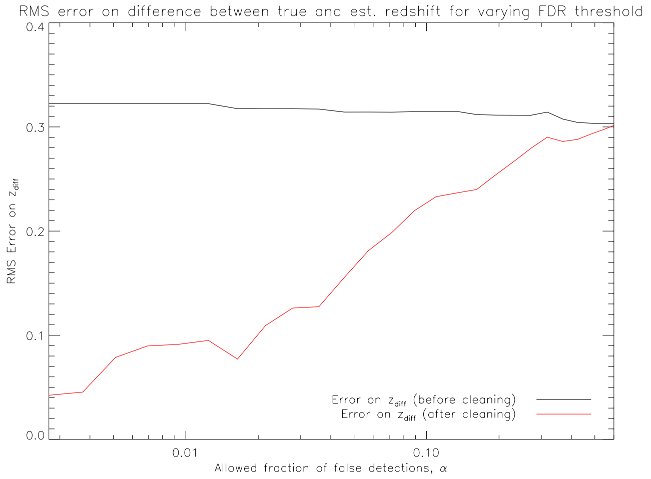


Fig. 9: In this figure, we demonstrate the effect of varying the FDR threshold on the RMS value of z_{diff}, for the mixed signal-to-noise catalogue.

5. Conclusions

As 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.

We can successfully make an estimate of the continuum without any modelling of galaxies and we can confidently make use of data at signal-to-noise levels that were previously beyond the reach of other techniques. This is achieved by denoising data with an appropriately chosen false detection rate threshold and implementing a simple peak-counting criterion, resulting in very high purity redshift estimates for a subset of galaxies in the catalogue.

This is the most useful aspect of Darth Fader – it can be used as a flagging mechanism to segregate what is likely to be good data for redshift estimation from what is likely to be erroneous, with a good level of confidence. Even at signal-to-noise levels as low as 0.5 on the spectrum, we are able to attain a purity of 95%, whilst still retaining 60% of the data.

Darth Fader represents a potential greater reach of spectroscopic surveys in terms of depth, since the faintest (and thus noisiest) galaxies in a survey – those at the detection limit of the instrument – will be primarily of higher redshifts. Current methods of spectroscopic survey redshift estimation usually employ signal-to-noise cuts or magnitude/flux limits, resulting in low signal-to-noise data being treated as unusable. Darth Fader offers a reliable and fully automated classification that works even down to very low signal to noise, with a purity and completeness that can be well understood through tests on realistic simulations.

There are, however, areas where the Darth Fader algorithm, as well as the catalogues used for testing, could be improved.

The spectra in these catalogues are, of course, simulated, and as such do not express the full complexity of real galactic spectra. Furthermore, both the template and galaxy catalogues used in this paper, consist primarily of strong emission line spectra, with some weak emission line spectra also included. The catalogue would benefit from the inclusion of a more representative sample including a more varied population of galaxy types. The Darth Fader algorithm has a requirement for the galactic spectra undergoing redshift estimation, to be wide enough in their wavelength range to encompass *at least* three main features. Spectra that are too short cannot currently be classified under the three peak-counting criterion.

The galaxy spectra are constructed to have a flat, wavelength independent, gaussian error across the whole spectrum. This is not realistic since we expect that real errors could be wavelength/pixel dependent, and potentially non-gaussian. This presents a greater challenge for denoising, however, the star1d routine used in Darth Fader has numerous optional parameters to deal with non-gaussian noise that could be applicable to such a task.

Continuum removal with wavelets, when pitted against elaborate modelling, may be seen as a comparatively 'coarse' method; however, there is no loss of generality in its usage in cross-correlation based redshift estimation methods, and it benefits from being a blind method requiring no prior knowledge of how galactic spectra arise.

We expect that the Darth Fader algorithm can be improved by utilising photometric information in a prior for the two-peak cases – which are a significant source of completeness degradation. The extra information provided by photometry should be sufficient to distinguish which line-pairs are presented on (more coarsely resolved) spectra, and hence resolve the previous 'confusion' between different line-pairs being mistaken for one another. It should be evident that Darth Fader can then be applied to narrower wavelength ranged spectra than previously, since only two main features would be required for positive redshift estimation instead of three.

The wavelet-based continuum-subtraction procedure used in Darth Fader is in principle not limited to galactic spectra, and preliminary tests suggest that it will prove useful for the continuum-modelling of the more structurally rich spectra of stars.

Darth Fader is clearly useful for both redshift estimation and empirical continuum estimation and will be made publicly available as part of the **iSAP** suite of codes.

Acknowledgements. The authors would like to acknowledge the support provided by the European Research Council, through grant SparseAstro (ERC-228261), FBA acknowledges support from the Royal Society via an URF. We also wish to thank Stephane Arnouts for his assistance with the LePhare program, and we also thank Jérémy Rapin & Simon Beckouche for their helpful feedback and assistance in the data presentation.

References

Aihara, H., Allende Prieto, C., An, D., et al. 2011, ApJS, 193, 29
Arnouts, S., Cristiani, S., Moscardini, L., et al. 1999, Monthly Notices of the Royal Astronomical Society, 310, 540
Bruzual, G. & Charlot, S. 2003, MNRAS, 344, 1000
Coleman, G. D., Wu, C., & Weedman, D. W. 1980, Astrophysical Journal Supplement Series, 43, 393
Costero, R. & Osterbrock, D. E. 1977, ApJ, 211, 675
Fligge, M. & Solanki, S. K. 1997, Astronomy & Astrophysics Supplement Series, 124, 579
Garilli, B., Fumana, M., Franzetti, P., et al. 2010, Publications of the Astronomical Society of the Pacific, 122, 827
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
Kinney, A. L., Calzetti, D., Bohlin, R. C., et al. 1996, Astrophysical Journal, 467, 38
Koski, A. T. & Osterbrock, D. E. 1976, ApJ, 203, L49
Kurtz, M. J. & Mink, D. J. 1998, Publications of the Astronomical Society of the Pacific, 110, 934
Lutz, R., Schuh, S., Silvotti, R., et al. 2008, in Astronomical Society of the Pacific Conference Series, Vol. 392, Hot Subdwarf Stars and Related Objects, ed. U. Heber, C. S. Jeffery, & R. Napiwotzki, 339
Panuzzo, P., Vega, O., Bressan, A., et al. 2007, Astrophysical Journal, 656, 206
Starck, J.-L., Bijaoui, A., & Murtagh, F. 1995, CVGIP: Graphical Models and Image Processing, 57, 420–431
Starck, J.-L. & Murtagh, F. 1994, Astronomy & Astrophysics, 288, 342
Starck, J.-L. & Murtagh, F. 2006, Astronomical Image and Data Analysis (Springer), 2nd edn.
Starck, J.-L., Murtagh, F., & Fadili, M. 2010, Sparse Image and Signal Processing (Cambridge University Press)
Stoughton, C., Lupton, R. H., Bernardi, M., et al. 2002, Astronomical Journal, 123, 485
SubbaRao, M., Frieman, J., Bernardi, M., et al. 2002, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 4847, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, ed. J.-L. Starck & F. D. Murtagh, 452–460
Tonry, J. & Davis, M. 1979, Astronomical Journal, 84, 1511
Yamada, I. 2001, in Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications, ed. D. Butnariu, Y. Censor, & S. Reich (Elsevier)