



# Darth Fader: Using wavelets to obtain accurate redshifts of spectra at very low signal-to-noise

D. P. Machado, A. Leonard, J.-L. Starck, F. B. Abdalla, S. Jouvel

## ► To cite this version:

D. P. Machado, A. Leonard, J.-L. Starck, F. B. Abdalla, S. Jouvel. Darth Fader: Using wavelets to obtain accurate redshifts of spectra at very low signal-to-noise. *Astronomy and Astrophysics* - A&A, EDP Sciences, 2013, 560, pp.A83. <10.1051/0004-6361/201219857>. <cea-01135491>

**HAL Id: cea-01135491**

**<https://hal-cea.archives-ouvertes.fr/cea-01135491>**

Submitted on 25 Mar 2015

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Darth Fader: Using wavelets to obtain accurate redshifts of spectra at very low signal-to-noise

D. P. Machado<sup>1</sup>, A. Leonard<sup>1</sup>, J.-L. Starck<sup>1</sup>, F. B. Abdalla<sup>2</sup>, and S. Jovel<sup>2,3</sup>

<sup>1</sup> CEA Saclay, IRFU, Service d'Astrophysique, Bt 709 – Orme des Merisiers, 91191 Gif-Sur-Yvette Cedex, France  
e-mail: daniel.machado@cea.fr

<sup>2</sup> University College London, Department of Physics & Astronomy, Kathleen Lonsdale Building, Gower Place, WC1E 6BT London, UK

<sup>3</sup> Institut de Ciències de l'Espai (IEEC-CSIC), 08193 Bellaterra, Barcelona, Spain

Received 20 June 2012 / Accepted 13 September 2013

## ABSTRACT

**Context.** Accurate determination of the redshifts of galaxies comes from the identification of key lines in their spectra. Large sky surveys, and the sheer volume of data they produce, have made it necessary to tackle this identification problem 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 (denoised and automatic redshifts thresholded with a false detection rate), which is a new wavelet-based method for estimating redshifts of galaxy spectra. Automated, simple, and largely empirical, we highlight how the Darth Fader algorithm performs in the very low signal-to-noise regime, and demonstrate its effectiveness at removing catastrophic failures from the catalogue of redshift estimates.

**Methods.** We present a new, nonparametric method for estimating and removing the continuum in noisy data that requires no a priori information about the galaxy properties. This method employs wavelet filtering based on a tuneable false detection rate (FDR) threshold, which effectively removes the noise in the spectrum, and extracts features at different scales. After removal of the continuum, the galaxy spectra are then cross-correlated with the eigentemplates, and a standard  $\chi^2$ -minimisation used to determine the redshift of the spectrum. FDR filtering is applied to the spectra a second time to determine the number of spectral features in each galaxy spectrum, and those with fewer than six total features are removed from the catalogue as we are unlikely to obtain a reliable and correct estimate of the redshift of such spectra.

**Results.** Applying our wavelet-based cleaning algorithm on a simulated testing set, we can successfully build a clean catalogue including extremely low signal-to-noise data ( $S/N = 2.0$ ), for which we are able to obtain a 5.1% catastrophic failure rate in the redshift estimates (compared with 34.5% prior to cleaning). We also show that for a catalogue with uniformly mixed signal-to-noise ratios between 1.0 and 20.0, with realistic pixel-dependent noise, it is possible to obtain redshifts with a catastrophic failure rate of 3.3% after cleaning (as compared to 22.7% before cleaning). Whilst we do not test this algorithm exhaustively on real data, we present a proof of concept of the applicability of this method to real data, showing that the wavelet filtering techniques perform well when applied to some typical spectra from the Sloan Digital Sky Survey archive.

**Conclusions.** The Darth Fader algorithm provides a robust method for extracting spectral features from very noisy spectra: spectra for which a reliable redshift cannot be measured are automatically identified and removed from the input data set. The resulting clean catalogue, although restricted in number, gives an extremely low rate of catastrophic failures, even when the spectra have a very low S/N. For very large sky surveys, this technique may offer a significant boost in the number of faint galaxies with accurately determined redshifts.

**Key words.** methods: data analysis – techniques: spectroscopic – galaxies: distances and redshifts – surveys

## 1. Introduction

The simplest method for estimating the redshift of a galaxy spectrum is by visual inspection, however, large sky surveys are providing astronomers with increasingly large datasets. The sheer number of spectra being obtained in these surveys make it necessary to make use of automated algorithms to obtain accurate information, as well as sophisticated techniques for dealing with the presence of noise; something which is increasingly important for distant and 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  $H_\alpha$  and the [O III] doublet, to their respective redshifted counterparts. When faced with just a single emission line, assuming it to be  $H_\alpha$  or [O II] 3727 Å is a viable option for spectroscopic redshift determination of emission line galaxies (since one of these is usually, but not always, the strongest feature; Le Fèvre et al. 1995, 2005). In such cases, degeneracies on the lines may potentially be resolved with the inclusion of photometric data. This type of approach is used in the SDSS Early Data Release (Stoughton et al. 2002).

Template matching methods for redshift estimation involve a “test set” – a catalogue of galaxy spectra with unknown redshifts – being matched to a set of template spectra initially at zero

redshift (“a template set”). The template spectra are matched to the data, typically using a  $\chi^2$  test or maximum likelihood estimator to determine the shift in wavelength between the template and the galaxy spectrum, and hence the redshift of that galaxy. Templates may come from simulated or semi-empirical spectra based on sophisticated modelling, local galaxy spectra whose redshifts are small and precisely known, or from a subset of high signal-to-noise spectra within the survey itself with redshifts that can be confidently identified.

Cross-correlation methods such as that described by Glazebrook et al. (1998) use a discrete Fourier transform (DFT) to correlate a template spectrum with a galaxy spectrum allowing the shift of the template spectrum (and thus redshift) to become a free parameter. Cross-correlation methods are convenient because they can be computed as a simple multiplication in Fourier space between the template and galaxy spectra, resulting in easier and faster computation than performing the same procedure in real space. These kinds of cross-correlation methods, however, require the spectra to be free of continuum in order to correctly correlate line features, with the presence of lines being a stronger constraint for redshift estimation than the shape of the continuum. Failure to subtract the continuum could result in erroneous cross-correlations since spectra with similar continua – but different lines – can give stronger correlations than those with different continua but similar lines.

Currently, continua have to either be modelled from population synthesis models (Bruzual & Charlot 2003; Panuzzo et al. 2007), requiring a priori knowledge of galactic properties and physics, or they are computed from feature-deficient galactic spectra of a similar continuum type, 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). Polynomial-fitting/statistical averaging 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 signal-to-noise (S/N) limit, it becomes exceedingly difficult to pinpoint exactly where the continuum lies, and polynomial fitting is not ideal.

In this paper, we introduce a new 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 at low S/N. We demonstrate on simulated data that this method performs well in both the low and high S/N regimes. Using a standard cross-correlation method for estimating the redshifts of galaxies, we demonstrate that an additional wavelet filtering to extract important features in each spectrum allows us to derive a selection criterion for galaxies for which we are able to accurately determine a redshift. Together this allows the DARTH FADER (denoised and automatic redshifts thresholded with a false detection rate) algorithm to effectively clean our galaxy catalogue by removing catastrophic failures, resulting in a galaxy redshift catalogue with a guaranteed high success rate, and allowing us to accurately and confidently determine the redshifts of galaxies even in the very low S/N regime. This will be useful in large photometric surveys, such as the upcoming Euclid survey (Refregier et al. 2010; Laureijs et al. 2011), as the calibration of photometric redshifts requires spectroscopic redshift information with a very low incidence of catastrophic failures.

This paper is organised as follows. In Sect. 2, we describe in detail the cross-correlation method used for redshift estimation. In Sect. 3, we briefly describe the construction of our datasets. In Sect. 4, we describe our wavelet-based continuum subtraction method, and the selection criteria for cleaning the catalogue.

In Sect. 5, we show results for various S/Ns, using wavelength-independent white-Gaussian noise, and a mixed S/N catalogue with non-stationary (pixel dependent) Gaussian noise, and compare the proportion of catastrophic failures in cleaned catalogues to that obtained using the full spectral catalogue. In Sect. 6, we demonstrate the potential applicability of this software to real data. We demonstrate successful feature extraction on several real spectra obtained from the SDSS data archive. Lastly, in Sect. 7 we summarise the key features of the DARTH FADER algorithm, and identify potential future applications of the algorithm and the methods involved.

## 2. Redshift estimation by cross-correlation

To estimate galaxy redshifts, we employ a cross-correlation method similar to that described by Glazebrook et al. (1998). This method involves a cross-correlation of test galaxy spectra at unknown redshift with template spectra.

We assume that any test spectrum  $S'_\lambda$  may be represented as a linear combination of template spectra  $T_{i\lambda}$ ,

$$S'_\lambda = \sum_i a_i T_{i\lambda}, \quad (1)$$

where each template spectrum is normalised according to

$$\sum_\lambda T_\lambda^2 = 1. \quad (2)$$

If we choose to bin our spectra on a logarithmic wavelength axis, redshifting becomes proportional to a translation,

$$\begin{aligned} \Delta &= \log(1+z) \\ &= \log(\lambda_{\text{observed}}) - \log(\lambda_{\text{rest frame}}). \end{aligned} \quad (3)$$

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

$$\chi^2(\Delta) = \sum_\lambda \frac{w_\lambda^2}{\sigma_\lambda^2} \left[ S_\lambda - \sum_i a_i(\Delta) T_{i(\lambda+\Delta)} \right]^2. \quad (4)$$

We can obtain the values of the expansion coefficients,  $a_i$ , by maximising Eq. (4) with respect to  $a_i$ . Following the prescription in Glazebrook et al. (1998), we take the weighting function,  $w_\lambda$ , and the normally distributed errors,  $\sigma_\lambda$ , to be wavelength independent and constant, which gives

$$a_i(\Delta) = \frac{\sum_\lambda S_\lambda T_{i(\lambda+\Delta)}}{\sum_\lambda T_{i(\lambda+\Delta)}^2}. \quad (5)$$

The numerator in Eq. (5) is simply the cross-correlation of the galaxy spectrum with the  $i$ th template spectrum. Substituting back into Eq. (4), we obtain

$$\chi^2(\Delta) \propto \sum_\lambda \left[ S_\lambda^2 - \sum_i a_i^2(\Delta) T_{i(\lambda+\Delta)}^2 \right]. \quad (6)$$

For a large test catalogue that includes a variety of galaxy types, a large number of templates is needed to ensure the best match-up between template and test spectra. To use all of them in the cross-correlation would be excessively time-consuming. If it

were possible to reduce the number of templates whilst still retaining most of the information content of these templates then we can render the method more practical.

Principal component analysis (PCA) is a simple tool that allows us to do just that: to reduce the dimensionality of this problem by extracting the most important features from our set of template spectra, 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 simpler to manipulate:

$$C_{ij} = \sum_{\lambda} T_{i\lambda} T_{j\lambda}^T. \quad (7)$$

Since this correlation matrix is always real and square-symmetric, it follows that it can be diagonalised,

$$\mathbf{C} = \mathbf{R}\mathbf{A}\mathbf{R}^T, \quad (8)$$

where  $\mathbf{A}$  represents the matrix of ordered eigenvalues (largest to smallest) and  $\mathbf{R}$ , the matrix of correspondingly ordered eigenvectors. The eigentemplates,  $\mathbf{E}$ , can then be obtained:

$$E_{j\lambda} = \frac{\sum_i R_{ij}^T T_{i\lambda}}{\sqrt{\Lambda_j}}, \quad (9)$$

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

$$\sum_{\lambda} E_{i\lambda} E_{j\lambda}^T = \delta_{ij}. \quad (10)$$

The effect of PCA is that it re-orientates the dataset to lie along the orthogonal eigenvectors (axes) sorted by descending variance. It effectively creates an ‘‘importance order’’, such that the eigenvector with the greatest variance (largest eigenvalue) will tend to correspond to the 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. For example, if  $H_{\alpha}$  is a very prominent feature in most of the template spectra, it will be present in one or more of the first few eigentemplates.

With this in mind we can now re-cast Eq. (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:

$$S_{\lambda} \simeq \sum_{i=1}^N b_i(\Delta) E_{i(\lambda+\Delta)}, \quad (11)$$

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

Using the orthogonality condition from Eq. (10), Eqs. (5) and (6) then become

$$b(\Delta) = \sum_{\lambda} S_{\lambda} E_{i(\lambda+\Delta)}, \quad (12)$$

$$\chi^2(\Delta) \propto \sum_{\lambda} S_{\lambda}^2 - \sum_{i=1}^N b_i^2(\Delta). \quad (13)$$

We then observe that the first term in Eq. (13) is a constant in the  $\chi^2$  function, and can be disregarded; therefore minimising

the  $\chi^2$  function in Eq. (13) is equivalent to *maximising* the related function,  $\tilde{\chi}^2$ , defined as

$$\chi^2(\Delta) \sim \tilde{\chi}^2(\Delta) = \sum_{i=1}^N b_i^2(\Delta). \quad (14)$$

Hence,  $\tilde{\chi}^2(\Delta)$  is computed by first computing the cross-correlation of each of the  $N$  retained eigentemplates  $E_i$  with the galaxy spectrum (Eq. (12)), and then summing  $b_i^2(\Delta)$  over these eigentemplates. 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 advantage that  $\Delta$  becomes a free parameter. Hence we obtain

$$b_i(\Delta) = \mathcal{F}^{-1}(\hat{S}_k \hat{E}_{ik}) = \frac{1}{M} \sum_{k=0}^{M-1} \hat{S}_k \hat{E}_{ik} e^{\frac{2\pi i k \Delta}{M}}, \quad (15)$$

and

$$\tilde{\chi}^2(\Delta) = \sum_{i=1}^N \left[ \mathcal{F}^{-1}(\hat{S}_k \hat{E}_{ik}) \right]^2, \quad (16)$$

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

Now that we have obtained Eq. (16) it is an easy task to extract the estimate for the redshift,  $z$ . The  $\tilde{\chi}^2$  function reaches a maximum 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  $\Delta = \Delta_{\tilde{\chi}^2} (= \Delta|_{\tilde{\chi}^2 = \tilde{\chi}^2_{\max}})$ , giving

$$z_{\text{est}} = 10^{\delta_s \Delta_{\tilde{\chi}^2}} - 1, \quad (17)$$

where  $\delta_s$  is the grid spacing on the  $\log_{10}$ -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 from each spectrum that are being matched to one another, rather than to any continuum features, 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 Sect. 4.3.1.

### 3. Simulations

The redshift estimation method described above requires two separate spectral catalogues: a set of galaxy spectra with noise and covering a range of redshifts that we aim to estimate (the test catalogue), and a set of noise-free, zero-redshift template spectra. We use the ‘‘Cosmos Mock Catalogue’’ (CMC) set of simulations as provided by [Jouvel et al. \(2009\)](#); [Zoubian & Kneib \(in prep.\)](#), which are based on the observed COSMOS SEDs of [Ilbert et al. \(2009\)](#); [Capak \(2009\)](#). We then rebin a randomly selected sub-sample of the CMC master catalogue onto an evenly spaced  $\log_{10} \lambda$  wavelength grid, spanning the range between 3000 Å to 10 500 Å for the test catalogue, with a wider range for the template spectra of 3000 Å to 20 900 Å.

The template catalogue was compiled by randomly selecting galaxies within the simulation set with redshift less than  $z = 0.1$ . The number of galaxies selected was chosen to be roughly 10% of the size of the main test catalogue, in order to ensure a representative sample of template galaxies. This resulted in a template catalogue of 277 simulated spectra, which were then blueshifted to be at zero redshift, and a test catalogue of 2860 simulated spectra with redshifts in the range  $0.005 < z < 1.7$ .

This choice of binning, and a similar pixelisation scheme as in Smees et al. (2013), gives a constant resolution across the spectrum of  $R(=\lambda/\Delta\lambda) \sim 850$  for all the catalogues, and a grid spacing of  $\delta_s = 2.17 \times 10^{-4} \log_{10} \text{Å}$ ; as compared to SDSS where the resolution and grid spacing are  $R \sim 1845$ , and  $\delta_s = 1.0 \times 10^{-4} \log_{10} \text{Å}$  respectively.

The template spectra are required to have a larger wavelength span than the test spectra since they must be able to accommodate a large enough overlap to identify the correct (global) cross-correlation minima with these test spectra at all redshifts under consideration. A restricted wavelength span on the set of template spectra will necessarily reduce the maximum redshift at which you can cross-correlate. This frequently will result in the cross-correlation picking a local minimum that exists in the overlap – since the global minimum lies outside this overlap – often resulting in a confusion between one principal feature for another.

Wavelength-independent (white) Gaussian noise was then added to the test catalogue to generate several catalogues in which all the galaxies had the same S/N. We define our S/N in the same manner as in the SDSS pipeline (Bolton et al. 2012, for BOSS/SDSS III)<sup>1</sup>, relative to the median S/N in the SDSS  $r$ -band filter (Fukugita et al. 1996):

$$S/N_r = \text{median} \left[ \frac{\text{flux}}{\sigma} \right]_{5600 \text{Å}}^{6760 \text{Å}} \quad (18)$$

where  $\sigma$  is the standard deviation of our added white-Gaussian noise, and the subscript  $r$  denotes that the median is calculated between the bounds of SDSS  $r$ -band filter (5600 Å to 6760 Å).

We choose this particular definition of S/N so as to be consistent with a realistic survey such as SDSS. The specific choice of SDSS band on which to base the S/N definition does not affect the method presented in this paper, however the motivation for choosing the  $r$ -band over any of the other SDSS bands is fully explained in Strauss et al. (2002). Whilst this definition of S/N is a good proxy for S/N on the continuum, and as such allows a simple comparison between different spectra, it should be cautioned that it is not necessarily a good proxy for the S/N on specific features.

An additional mixed S/N catalogue was generated by adding pixel-dependent Gaussian noise such that the spectra in the catalogue had a uniform distribution in S/N in the range  $1.0 < S/N < 20$  as described in Sect. 6.

#### 4. The DARTH Fader algorithm

DARTH Fader works in a number of different steps to take a test catalogue of galaxy spectra containing both spectral lines and continuum features (as well as noise), and to output a clean catalogue of galaxies for which we are able to obtain robust and accurate redshift estimates via PCA and cross-correlation. A schematic of the full algorithm is shown in Fig. 1.

<sup>1</sup> The idlspec2d pipeline software package is available at: <http://www.sdss3.org/dr8/software/products.php>

In order to estimate the redshift of galaxies by cross-correlation, both the templates and the test galaxy spectra must be continuum-free. Current methods for continuum subtraction rely on a handful of principal techniques: careful modelling of the physics of galaxies to estimate the continuum, a matching of continua between featureless galaxy spectra (typically sourced from elliptical galaxies or galactic bulges) and the spectra from which we wish to remove the continuum, or a polynomial fitting. (Koski & Osterbrock 1976; Costero & Osterbrock 1977; Panuzzo et al. 2007). Median filtering to remove spectral features is also a possibility; however this relies on a fixed choice for the median filtering window size, which may not be appropriate for resolved lines or blended line doublets.

The first two of these methods have the disadvantage of requiring some knowledge of galaxy physics (which may not be precisely known), and being somewhat restricted to lower redshift/higher S/N galaxies. Careful modelling is computationally intensive and liable to result in failure if unusual galaxy types are found, or if the physics involved is not fully understood or modelled well enough. Continuum-matching methods require a priori knowledge of the galaxy type of a set of galaxies, and/or are reliant on the correspondence between similar looking continua (with one relatively featureless, in order to remove one from the other) which may not exist for all spectra. All matching methods have a problem at high noise levels where different but superficially similar spectra are mistakenly associated. Polynomial fitting is a further alternative that is limited to high signal-to-noise spectra, or spectra that have been denoised beforehand (as applied to the SDSS data release, Stoughton et al. 2002).

By contrast our new method of continuum subtraction is completely empirical, requires no knowledge of the physics or type of the galaxy involved and can be used even with very noisy spectra. This method relies on a multiscale modelling of the spectra, as described below.

##### 4.1. Spectra modelling

We model the galaxy spectrum as a sum of three components – continuum, noise and spectral lines:

$$S = L + N + C, \quad (19)$$

where  $L$  contains the spectral line information,  $N$  is the noise and  $C$  is the continuum.  $L$  can be decomposed into two parts, emission lines  $L_e$  and absorption lines  $L_a$ :  $L = L_e + L_a$ , where, provided the continuum has been removed,  $L_e > 0$  and  $L_a < 0$ .

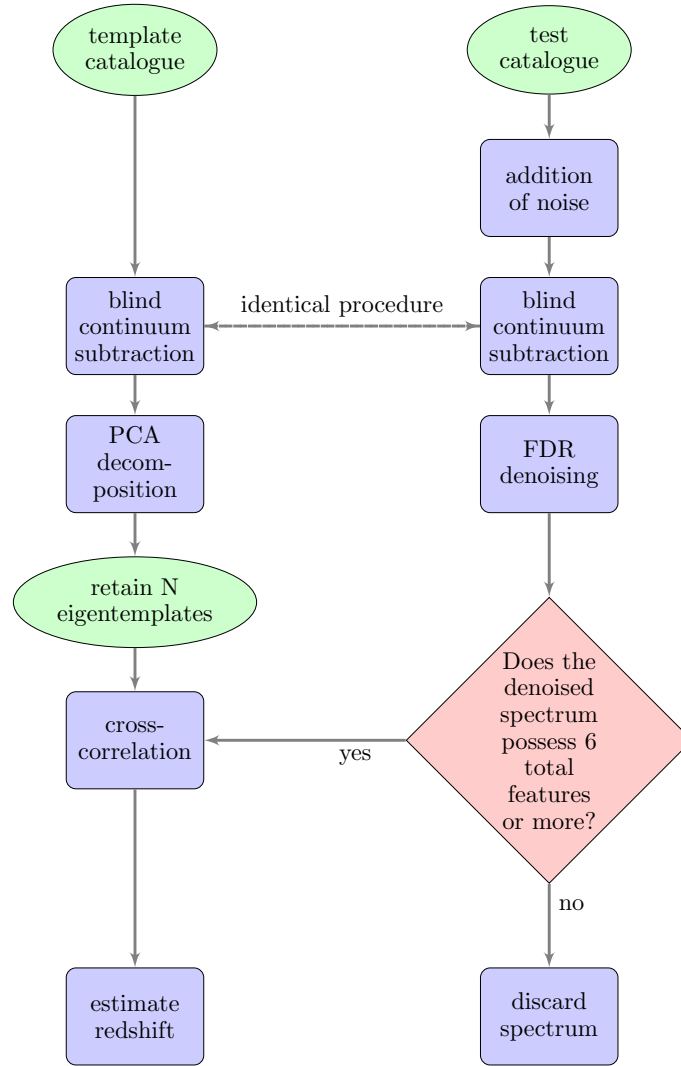
The problem is then to estimate these components,  $L$  and  $C$ , from a unique data set. This is possible assuming that features of  $L$  are important only on small and intermediate scales, while the continuum contains no small scale features, and is dominant on large scales only. However, several problems remain:

- Strong emission/absorption lines impact significantly at all frequencies, so a low pass filtering of the input is not sufficient to properly estimate the continuum,  $C$ .
- Both emission and absorption lines are difficult to detect because of the presence of noise.

##### 4.2. Continuum removal

The method we present in the following is done in four steps, two for continuum estimation and two for absorption and emission line estimation.

1. We first detect strong emission and absorption lines, which could be seen as outlier values for continuum emission.



**Fig. 1.** Operation of Darth Fader. The number of eigentemplates to be retained is at the discretion of the user, and may depend on the distribution of spectral types in the data. We have chosen to retain 20 eigentemplates because they encompass 99% of the total eigenvalue weight. In general the number retained will be significantly less than the number of spectra in the original template catalogue. The FDR denoising procedure denoises the positive and negative halves of the spectrum independently, with positivity and negativity constraints respectively. The requirement of six features or more in the denoised spectrum effectively cleans the catalogue of galaxies likely to yield catastrophic failures in their redshift estimates. It should be noted that a “no” decision represents the termination of that spectrum from the test catalogue and our analysis. Potentially, a further analysis could be employed at this point in order to review spectra possessing 5 features, for example, to see if further information, such as the association of standard lines to these features, can yield a redshift estimate (which could be used in conjunction with an estimate obtained from cross-correlation, with corresponding redshifts likely to be correct).

2. We subtract from the data these specific strong features and estimate the continuum from this.
3. We then re-estimate the emission lines from the original data now continuum-subtracted via steps 1 and 2.
4. We also re-estimate the absorption lines in a similar way.

#### 4.2.1. Strong line removal using the pyramidal median transform

In order to detect strong emission and absorption lines that could be seen as outliers for the continuum, we need a tool that is highly robust to these outliers. The choice of median filtering is generally the correct one for such a task. However, fixing the median filtering window size to the width of an unresolved line is not appropriate for blended line doublets and resolved lines. In our case, a better choice therefore is the multiscale median transform that was proposed for cosmic ray removal in infrared

data (Starck et al. 1996a; Starck & Murtagh 2006). Furthermore its pyramidal nature allows us to significantly speed up computation time (Starck et al. 1996b). In this framework, strong features of different width can be efficiently analysed.

In a general multiscale transform<sup>2</sup>, a spectrum of  $n$  bins,  $S_\lambda = S[1, \dots, n]$  can be decomposed into a coefficient set,  $W = \{w_1, \dots, w_J, c_J\}$ , as a superposition of the form

$$S_\lambda = c_J(\lambda) + \sum_{j=1}^J w_j(\lambda), \quad (20)$$

where  $c_J$  is a smoothed version of the original spectrum  $S_\lambda$ , and the  $w_j$  coefficients represent the details of  $S_\lambda$  at scale  $2^{-j}$ ; thus, the algorithm outputs  $J + 1$  sub-band arrays each of size  $n$ . The

<sup>2</sup> IDL routines to compute this and other wavelet transforms are included in the iSAP package available at: <http://www.cosmostat.org/software.html>

present indexing is such that  $j = 1$  corresponds to the finest scale or highest frequencies.

We use a similar multiscale transform in the following algorithm for strong line detection:

- take the pyramidal median transform (PMT) of the input spectrum  $S$  (a median window of size 5 was used in all our experiments), we get a set of bands  $w_j$  and  $c_j$  at different scales  $j$ ,  $\mathcal{P}(S) = \{w_1, \dots, w_J, c_J\}$ . Where  $w_j$  corresponds to multiscale median coefficients, and can be interpreted as the information lost between two resolutions when the downgrading is performed using the median filtering followed by a downsampling of factor 2. The  $c_j$  term corresponds to a very coarse resolution of the input signal. Full details can be found in [Starck et al. \(2010\)](#).
- for each band  $w_j$ , threshold all coefficients with an amplitude smaller than four times the noise level.
- set the coarse resolution,  $c_J$ , to zero.
- reconstruct the denoised spectrum  $S_1$ .

$S_1$  represents a crude estimation of the lines  $L$ , mainly because the noise behaviour in the pyramidal decomposition cannot be calculated as well as in a linear transform such as the Fourier or the wavelet transform. However the process is much more robust than with a linear transform since strong lines with small width will not contaminate the largest scales as it would be the case for instance with wavelets, resulting in artefacts termed “ringing”.

Since  $S_1$  contains the signal from strong lines,  $S_2 (= S - S_1)$  will be free of any strong features, and robust continuum estimation can easily be derived from it.

#### 4.2.2. Continuum extraction

The second step is therefore to estimate the continuum from  $S_2$ . The largest scale of  $S_2$  should contain the continuum information (see first term in Eq. (20)), whilst the noise and undetected lines are expected to be dominant on smaller scales. So now the coarsest scale in a wavelet decomposition, or any low pass filtering, would give us a good estimation for the continuum. The great advantage of wavelets for this task, as compared to a simple low pass filtering performed in Fourier space for example, is to allow a greater flexibility for handling the border (i.e. minimising edge effects), and there being no requirement to assume periodicity of the signal. We do this using the starlet wavelet transform, also called isotropic undecimated wavelet transform (Eq. (20)).

This transformation is simply a new representation of the original signal, which can be recovered through a simple summation. For a detailed description of the starlet transform see [Starck et al. \(2010\)](#), which has further been shown to be well-adapted to astronomical data where, to a good approximation, objects are commonly isotropic ([Starck & Murtagh 1994, 2006](#)).

We therefore estimate the continuum by first taking the wavelet transform of  $S_2$ , i.e.:  $W^{[S_2]} = \{w_1^{[S_2]}, \dots, w_J^{[S_2]}, c_J^{[S_2]}\}$ , and then retaining only the largest scale:  $c_J^{[S_2]} = C$ . This continuum can now be subtracted from the original noisy spectrum to yield a noisy, but now continuum-free, spectrum.

#### 4.2.3. Example

We show in Fig. 2 an example noise-free spectrum from our simulated catalogue, containing both line features and continuum. In Figs. 3a and 3b, we show the same spectrum with noise

added resulting in values of 5 and 1 for the SDSS  $r$ -band S/N. We note that galaxy surveys select galaxies based on their S/N, typically with a lower bound at an S/N of 5–10 on the continuum. Over-plotted in these latter two figures is the continuum as estimated by the method described above, for the spectrum with an S/N of 5, the continuum fit can be seen to be quite good. At lower S/N, the continuum fit is quite poor, as  $S_1$  is poorly estimated for this particular noise realisation, and the dominating influence of noise effectively conceals the continuum. However, this continuum estimate at low S/N is still well within the noise, and the correct order of magnitude. For reference we calculate the S/N on the  $H_\alpha$  line in each case by taking the ratio of the mean flux per pixel on the line and the noise standard deviation; the  $H_\alpha$  S/N was found to be 8.9, and 1.7 respectively for  $r$ -band S/Ns 5 and 1.

#### 4.3. Absorption/emission line estimation using sparsity

The wavelet representation of a signal is useful because it enables one to extract features at a range of different scales. In many cases, a wavelet basis is seen to yield a sparse representation of an astrophysical signal (such as a spectrum or a galaxy image), and this sparsity property can be used for many signal processing applications, such as denoising, deconvolution, and inpainting to recover missing data (e.g. [Fadili & Starck 2009](#); [Starck et al. 2010](#)). In this paper, we focus on one such application: that of denoising spectra using wavelet filtering.

The basic idea underlying sparse wavelet denoising is that the signal we are aiming to recover is sparsely represented in our chosen wavelet dictionary. This means that the signal is completely represented by a small number of coefficients in wavelet space<sup>3</sup>. This sparsity property means that if we are able to identify the important coefficients, it is straightforward to extract the signal from the noise.

There are various methods to do this; one simple method would be  $K\sigma$  clipping, where a threshold is set relative to an estimate of the noise, and all coefficients with an S/N less than  $K$  are set to zero. A more sophisticated method involves the use of a false detection rate (FDR) threshold, which allows us to control contamination from false positive lines arising from noise features. This method will be described in detail below. 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).

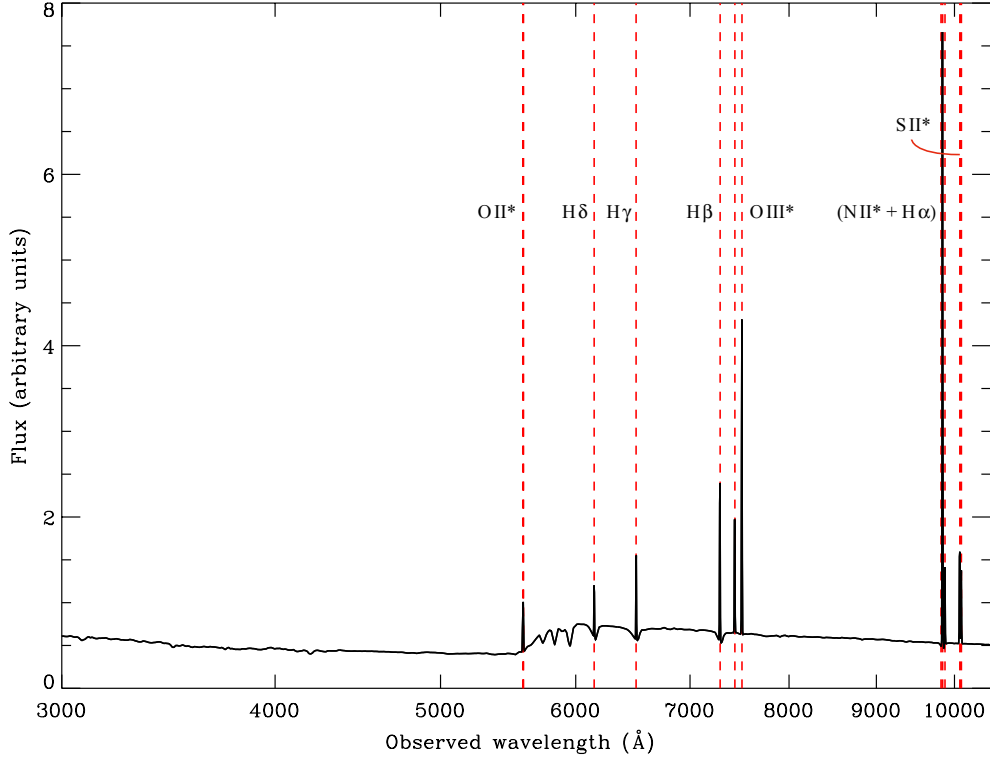
##### 4.3.1. Sparse Wavelet Modelling of spectra

As the continuum  $C$  is now estimated, we will now use the continuum free spectrum  $S_c = S - C$ . We can consider now that the remaining problem is to estimate the lines, assuming  $S_c = L + N$  and  $L = L_e + L_a$ . We exploit the wavelet framework in order to decompose it into two components: line features, and noise. This is done using a modified version of a denoising algorithm based on the hybrid steepest descent (HSD) minimisation algorithm developed by [Yamada \(2001\)](#).

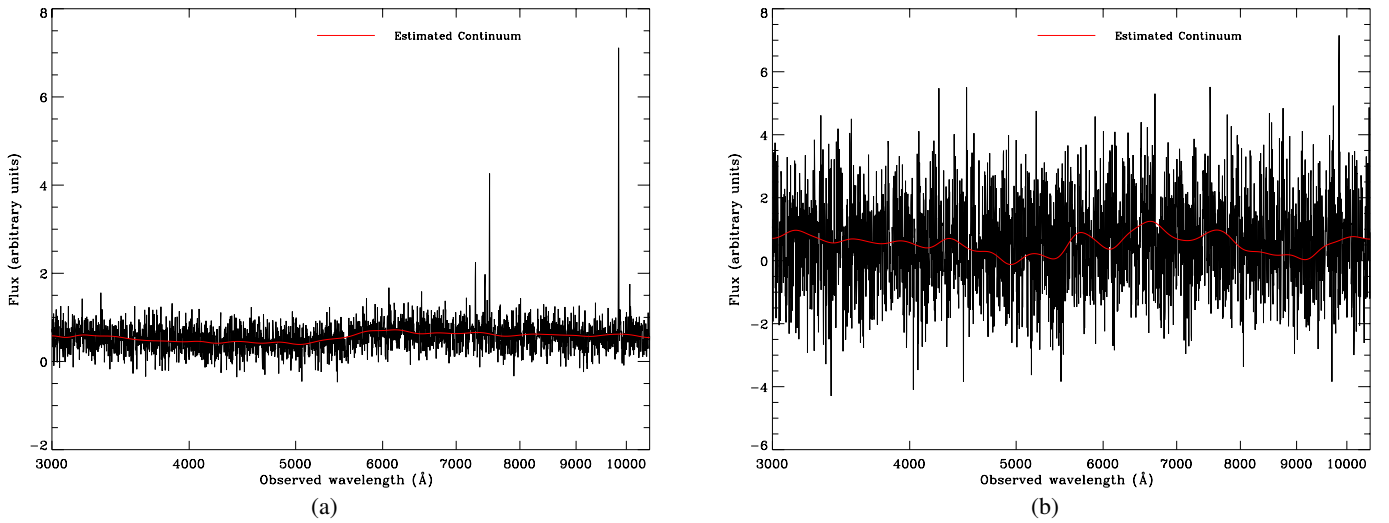
Hence, we can reconstruct  $L$  by solving the following optimisation problem:

$$\min_L \|\hat{W}L\|_1, \quad \text{s.t.} \quad S \in C, \quad (21)$$

<sup>3</sup> This is analogous to the representation of periodic signals in Fourier space, where they may be represented by only a few frequencies in this domain.



**Fig. 2.** Example spectrum from the test catalogue ( $z = 1.4992$ ), prior to the addition of noise. The main emission lines are labeled; with an asterisk denoting a doublet feature. The [O II] doublet is fully blended in this spectrum.



**Fig. 3.** Same as in Fig. 2 but with manually added white-Gaussian noise at a signal-to-noise level in the  $r$ -band of 5 in **a**), and of 1 in **b**). The red lines indicate the empirically-determined continua in each case. Many of the prominent lines are easily visible by eye at the higher S/N of 5, whereas at the lower S/N of 1 most of the lines are obscured, with only  $H_\alpha$  being sufficiently prominent so as to be detectable. The continuum estimate is good at the S/N of 5, and comparatively poor, but of the correct order of magnitude, at the lower S/N due to the dominating influence of noise. As an indication of line-S/N, we quote the values for the S/N on  $H_\alpha$  for these particular spectra as 8.9 and 1.7 respectively for **a**) and **b**).

where  $\hat{W}$  is the wavelet transform operator,  $\|\cdot\|_1$  is the  $\ell_1$  norm, which promotes sparsity in the wavelet domain, and  $C$  is convex set of constraints, the most important of which is a linear data fidelity constraint:

$$\left| w_j^{[S_c]}(\lambda) - w_j^{[L]}(\lambda) \right| \leq \varepsilon_j, \quad \forall (j, \lambda) \in \mathcal{M}. \quad (22)$$

Here  $w_j^{[S_c]}$  and  $w_j^{[L]}$  are respectively the wavelet coefficients of  $S_c$  and  $L$ , and  $\varepsilon_j$  is an arbitrarily small parameter that controls how

closely the solution  $L$  matches the input data. The constraint set  $C$  may also include further constraints, such as positivity for emission line-only spectra, etc. Note that the large scale coefficients  $c_J$  are not considered in this minimisation, as we do not expect the largest scales to contain any useful information since the continuum has been subtracted.  $\mathcal{M}$  is the *multiresolution support* (Starck et al. 1995), which is determined by the set of detected significant coefficients at each scale  $j$ , and wavelength  $\lambda$ ,



as

$$\mathcal{M} := \{(j, \lambda) \mid \text{if } w_j(\lambda) \text{ is declared significant}\}. \quad (23)$$

The multiresolution support is obtained from the noisy data  $S_c$  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  $\tau_j$ , often chosen as  $\tau_j = K\sigma_{j,\lambda}$ , where  $K$  is specified by the user (typically between 3 and 5) and  $\sigma_{j,\lambda}$  is the noise standard deviation at scale  $j$  and at wavelength  $\lambda$ . When the noise is white and Gaussian, we have  $\sigma_{j,\lambda} = \sigma_j$ , and  $\sigma_j$  can directly be derived from the noise standard deviation in the input data. When the noise is Gaussian, but not stationary, which is generally the case for spectral data, we can often get the noise standard deviation per pixel  $\sigma_\lambda$  from the calibration of the instrument used to make the observation, and  $\sigma_{j,\lambda}$  can be easily derived from  $\sigma_\lambda$  (Starck & Murtagh 2006).

An interesting and more efficient alternative to this standard  $K\sigma$  detection approach is the procedure to control the FDR. The FDR method (Benjamini & Hochberg 1995)<sup>4</sup> allows us to control the average fraction of false detections made over the total number of detections. It also offers an effective way to select an adaptive threshold,  $\alpha$ .

In the most general context, we wish to identify which pixels of our galaxy spectrum contain (predominantly) signal, and are therefore “active”, and those which contain noise and are therefore “inactive”. The measured flux in each pixel, however, may be attributed to either signal or noise, with each having an associated probability distribution. When deciding between these two competing hypotheses, the null hypothesis is that the pixel contains no signal, and the alternative hypothesis is that signal is present (in addition to noise).

The FDR is given by the ratio:

$$FDR = \frac{V_f}{V_a}, \quad (24)$$

where  $V_f$  is the number of pixels that are truly inactive (i.e. are part of the background/noise) but are declared to be active (falsely considered to be signal), and  $V_a$  is the total number of pixels declared active.

The procedure controlling the FDR specifies a fractional threshold,  $\alpha$ , between 0 and 1 and ensures that, *on average*, the FDR is no bigger than  $\alpha$ :

$$\langle FDR \rangle \leq \frac{V_i}{V_T} \cdot \alpha \leq \alpha. \quad (25)$$

The unknown factor  $V_i/V_T$  is the proportion of truly inactive pixels; where  $V_i$  is the number of inactive pixels, and  $V_T$  the total number of pixels.

A complete description of the FDR method can be found in Starck & Murtagh (2006) and, from an astrophysical perspective, in Miller et al. (2001). FDR has been shown to outperform standard methods for source detection (Hopkins et al. 2002), and Pires et al. (2006) have shown that FDR is very efficient for detecting significant wavelet coefficients for denoising of weak lensing convergence maps. In this paper, the FDR method is applied at each wavelet scale, and hence gives a detection threshold  $\tau_j$  per wavelet scale.

<sup>4</sup> Benjamini & Hochberg term FDR as false *discovery* rate in their paper; it is exactly analogous to what we term false detection rate in this paper.

The minimisation in Eq. (21) can be achieved using a version of the HSD algorithm adapted to non-smooth functionals, full details of which can be found in Starck et al. (2010).

In practice, we separately estimate emission lines and absorption lines by running the algorithm twice, first with a positivity constraint to get  $L_e$ , and then with a negativity constraint to estimate  $L_a$ . We found this approach more efficient than a single denoising without constraint, allowing us to reduce ringing (a type of denoising artefact that would compound feature counting) around detected lines. Our final estimate of  $L$  is then obtained by  $L = L_e + L_a$ .

#### 4.4. Example

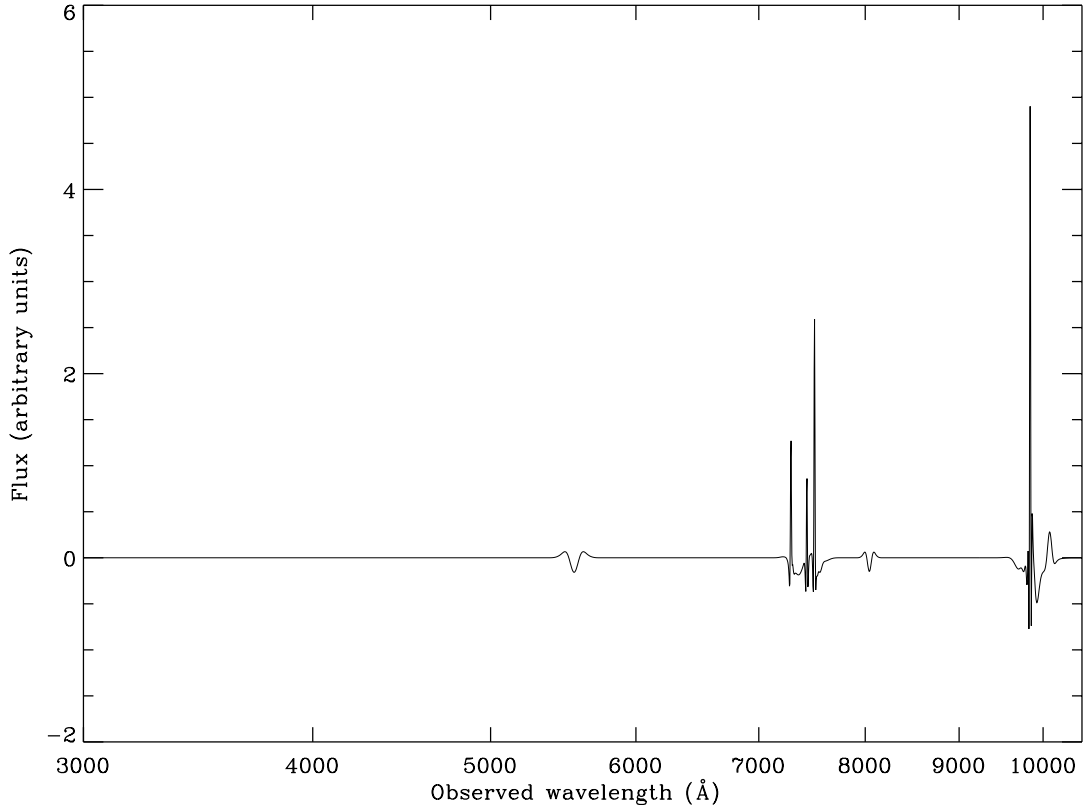
As an example, we show in Fig. 4 the first attempt at the reconstruction of the lines,  $L$ , from Fig. 3a, using an FDR threshold of  $\alpha = 4.55\%$ . Here, the positive and negative halves of the spectrum have *not* received independent treatment and the denoising is unrestricted since it is for the purpose of continuum-subtraction. It is the FDR denoising with the aim of feature-counting (Fig. 6) that requires a separate treatment of positive and negative halves of the spectrum; the FDR denoising in order to isolate the continuum does not require this procedure. The denoising of Fig. 3b fails to detect any features, and thus returns a null spectrum.

The secondary step – the denoising to determine the number of features – is shown in Fig. 6, for the continuum subtracted spectrum shown in 5a. Note how the denoising artefacts (termed ringing) in Fig. 4 are less present, and as such are not mis-counted as features. In the noisier example (Fig. 5b) the denoising once again fails to detect any features and returns a null spectrum (for this particular noise realisation and FDR threshold of 4.55% allowed false detections), and this would lead to the spectrum being discarded from our catalogue as unlikely to yield an accurate redshift estimate.

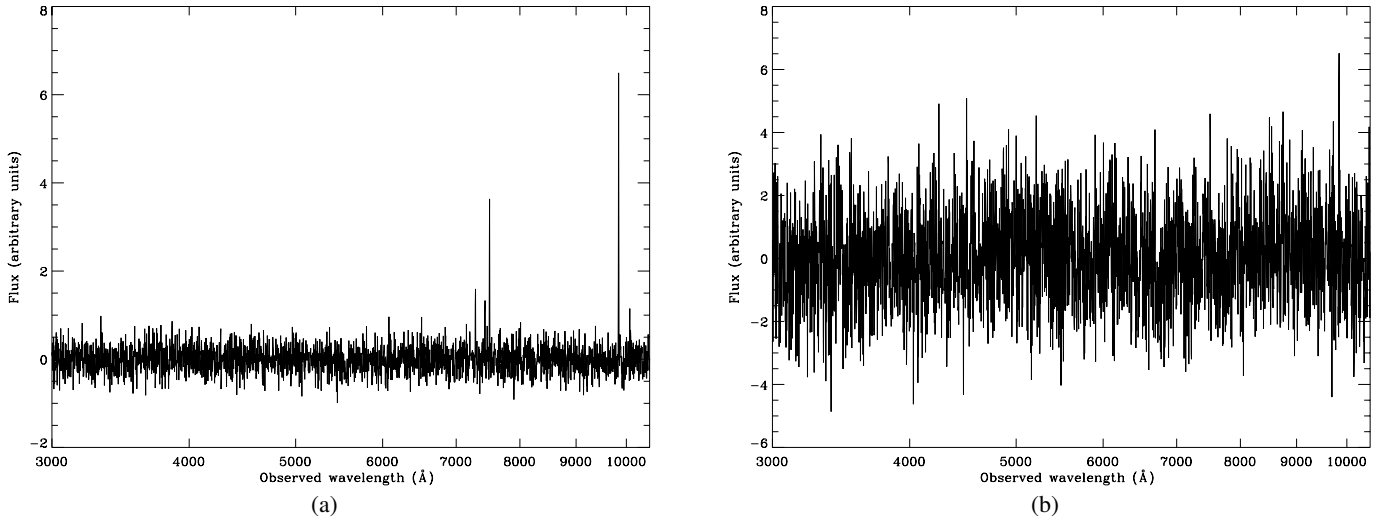
#### 4.5. Redshift estimation

For our method, we followed the PCA procedure as described in Sect. 2 on a set of noise-free template spectra to obtain eigentemplates, of which we kept only the first  $N$  principal components such that they comprised 99.93% of the total eigenvalue weight, which in our case resulted in the retention of 20 eigentemplates. We continuum-subtracted our test spectra as described in Sect. 4.3.1. Since the white-Gaussian noise on a spectrum will in principle be uncorrelated with the eigentemplates, we chose to use the noisy galaxy spectra in the cross-correlation. This ensured that we preserved all the line information in the spectrum, rather than discarding some of the signal through denoising, and hence we were not probing potential systematics of the denoising simultaneously with the redshift estimation.

However, when dealing with the pixel-dependent noise, it is the denoised spectra that must be used in the cross-correlation, since very noisy pixels in proximity to less noisy pixels will produce features that strongly resemble lines, and would thus be highly correlated with features in the eigentemplates (independent of the redshift of the spectra involved) if not denoised. For example, an error-curve peaking strongly at 7600 Å, may frequently produce features at this wavelength in the noisy spectra that strongly resemble lines. The effect of this false feature is to bias the cross-correlations such that large features in the templates (for example  $H_\alpha$ ) consistently match up to this false line, independent of the true redshift of the spectrum, resulting in redshift estimates that consistently favour an



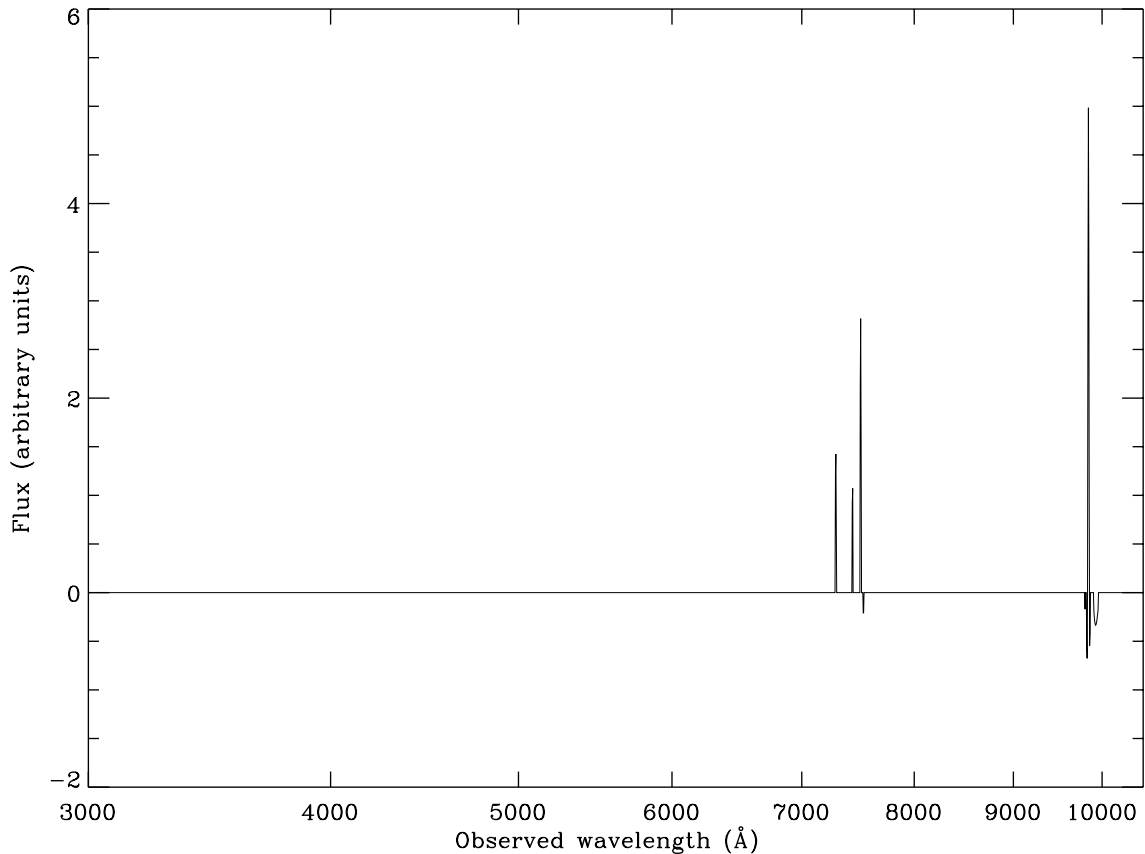
**Fig. 4.** Result of an unrestricted denoising of the spectrum in Fig. 3a with an FDR threshold corresponding to an allowed rate of false detections of  $\alpha = 4.55\%$ . The [O III] doublet,  $H_\alpha$  and  $H_\beta$  are all cleanly identified. There are small features corresponding to [O II] and [S II], and a spurious feature at just over 8000 Å. The FDR denoising of Fig. 3b fails to detect any features for this particular spectrum, noise-realisation and choice of FDR threshold, and thus returns a null spectrum (not shown).



**Fig. 5.** Panels (a) and (b) are the spectra as shown in Figs. 3a and 3b with their empirically determined continua subtracted.

incorrect redshift. In this example many spectra would be biased to have an estimated redshift of 0.158, irrespective of their true redshift values. As such we must use the denoised versions of the spectra with non-stationary noise for the cross-correlations. However, the redshift estimation will thus incur any potential systematics of the denoising procedure itself, this is explored further in Sect. 6.1.

Clearly, at low S/N, some of these cross-correlations will produce inaccurate results due to many features becoming lost in the noise. Higher S/N is not a guarantee of a successful redshift estimate; it is possible that line confusion, a lack of features, or poor representation in the basis of eigentemplates will result in a catastrophic failure.



**Fig. 6.** Result of denoising the positive and negative sections (shown together) of the spectrum shown in Fig. 5a with positivity and negativity constraints respectively. Note the reduced ringing, which leads to a more representative result with respect to the number of true features. Once again the FDR denoising of our noisier example (Fig. 5b) yields a null spectrum (not shown), and would thus result in the discarding of this spectrum from the redshift analysis.

A simple, but effective, criterion for the selection of galaxy spectra that will be likely to yield an accurate redshift estimate can be developed by considering the number of significant line features (either absorption or emission) present in the spectrum. For a spectrum containing many prominent features, it should be very easy to determine the redshift via cross-correlation with a representative set of eigentemplates. In cases where only one prominent feature is present, for example, we expect that the cross-correlation function will have several local maxima, each occurring when the location of the line feature in the test spectrum aligns with any of the features present in the (shifted) template spectrum. A similar effect would be expected for a spectrum with many – but not particularly prominent – features, obscured by noise. In such cases, it will not generally be possible to obtain a correct redshift unless we have more information about that feature or the spectrum (for example identifying the continuum shape/using photometric data which would help in identifying the colour of the galaxy; redder being indicative – but not definitively – of higher redshift), and/or we make an assumption that the most prominent feature is a specific standard line (for example,  $H_\alpha$ ). There is also the possibility that the dominant feature in the spectrum is a noise feature (this could be the case for multiple features if the spectrum is very noisy), in which case it will be impossible to estimate the redshift correctly.

With an increasing number of detected features, and a high degree of certainty that they are not the result of noise contamination, it should become clear that the redshift estimate obtained for such a test spectrum becomes progressively more reliable.

A question arises as to quite how many features are sufficient to distinguish reliable redshifts from those which are not reliable and we wish to discard. Through empirical tests, we have chosen 6 features in total as the criterion by which we decide the reliability of the redshift estimate of a test spectrum in our catalogue.

With this in mind, we use the denoising procedure described in Sect. 4.3.1 on the continuum-subtracted spectrum and identify the number of features present in the denoised spectrum via a simple feature-counting algorithm<sup>5</sup>. We then partition the catalogue in two: a cleaned catalogue comprised of noisy spectra for which denoising presents 6 or more features, where we keep the redshift determination as likely to be accurate; and a discarded catalogue with spectra only possessing 5 features or fewer upon denoising, where the redshift estimates are deemed to be unreliable.

Features are considered to be “peaks” 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* features. Employing this method alone would ignore absorption features; to overcome this we denoise and feature-count the positive and negative halves of the spectrum separately, independently detecting both emission and absorption features.

<sup>5</sup> Algorithm adapted from “peaks.pro”, available from: <http://astro.berkeley.edu/~johnjohn/idlprocs/peaks.pro>

At low S/N there is a trade-off between relaxing the FDR threshold to pick up more features – or indeed any features – and imposing a stricter threshold to prevent the detection of spurious lines. Recall that the FDR parameter constrains the average ratio of false detections to total detections. Therefore, for an FDR parameter of  $\alpha = 0.05$ , for example, we allow on average one false feature for every 20 features detected; i.e. an average ratio of 19 true features to 1 false feature. In very noisy data, it might not be possible to achieve this statistical accuracy, and therefore no features will be identified by the algorithm.

It follows that even if 6 features can be obtained from the denoising of the spectrum, some of them may still be spurious, and this could lead to an erroneous redshift estimate from cross-correlation (particularly if the spurious line is dominant, with this strongly biasing the cross-correlation) and false-positive contamination of our retained data. However, as noted, a maximum for this false line contamination is set by the FDR threshold,  $\alpha$ . In addition, the spectra that possess fewer than 6 features may provide redshift estimates that would otherwise be reliable; the criterion chosen leads them to be discarded. There exists this necessary trade-off between the fraction of catastrophic failures in the resulting redshift catalogue and the amount of data that is discarded.

## 5. Experimental results

In order to test our algorithm, we investigate the effect of the choice of the FDR parameter on the rate of catastrophic failures and the fraction of retained data in the cleaned catalogue at different S/N values. We use the simulated data described in Sect. 3, and apply the Darth Fader algorithm over multiple FDR thresholds, keeping the signal-to-noise constant; and again over catalogues with different S/Ns, keeping the FDR threshold constant. Lastly we apply Darth Fader to a uniformly mixed S/N catalogue with pixel-dependent Gaussian noise with S/N ranging from 1 to 20, utilising a range of values for the FDR threshold.

We define the retention  $\mathcal{R}$ ; catastrophic failure rates before cleaning,  $\mathcal{F}$ , and after cleaning,  $\mathcal{F}_c$ ; and capture rate  $C$  of the sample to be:

$$\mathcal{R} = \frac{\mathcal{T}_c}{\mathcal{T}} \times 100\%, \quad (26)$$

$$\mathcal{F}_c = \left(1 - \frac{\mathcal{U}_c}{\mathcal{T}_c}\right) \times 100\%, \quad (27)$$

$$C = \frac{\mathcal{U}_c}{\mathcal{U}} \times 100\%, \quad (28)$$

where  $\mathcal{T}$  and  $\mathcal{T}_c$  respectively denote the total number of galaxies in the sample (before cleaning) and the retained number of galaxies in the sample after cleaning (the number that satisfy the feature-counting criterion). Similarly,  $\mathcal{U}$  and  $\mathcal{U}_c$ , respectively denote the number of successful redshift estimates in the sample before and after cleaning. In Eq. (27), the brackets denote the option of calculating the catastrophic failure rate before cleaning (ignoring the subscripts) or the catastrophic failure rate after cleaning (inserting the subscript c everywhere shown). The number of successes after cleaning,  $\mathcal{U}_c$ , cannot be greater than  $\mathcal{U}$ , hence the capture rate represents the proportion of correct estimates available before cleaning that are retained post-cleaning.

We present the result of cleaning the catalogue using an FDR threshold of  $\alpha = 4.55\%$  on a catalogue of spectra with an S/N of 2.0 in Fig. 7. The two panels compare the distribution of redshift estimates before and after cleaning of the catalogue using

the feature-counting criterion. A clear improvement is seen when cleaning is applied: the fraction of catastrophic failures in the catalogue is reduced from 34.5% before cleaning to 5.1% after cleaning. In addition, we have retained 76.2% of the galaxies which yielded a correct redshift estimate before cleaning (the capture rate), with the retained catalogue comprising 52.6% of the total number of galaxies in the test catalogue.

Prior to cleaning there clearly exist two components to the redshift estimates, a strong square diagonal (the  $x = y$  line where the redshift estimates are likely correct) and a cloud of misidentifications (with a small, non-square, diagonal component) where the estimated redshifts are generally underestimates of the true redshift. It is important to note that failures at this point are due to the standard cross-correlation, with non-square diagonal components often being indicative of line confusion (for example between  $H_\alpha$  and [O III]).

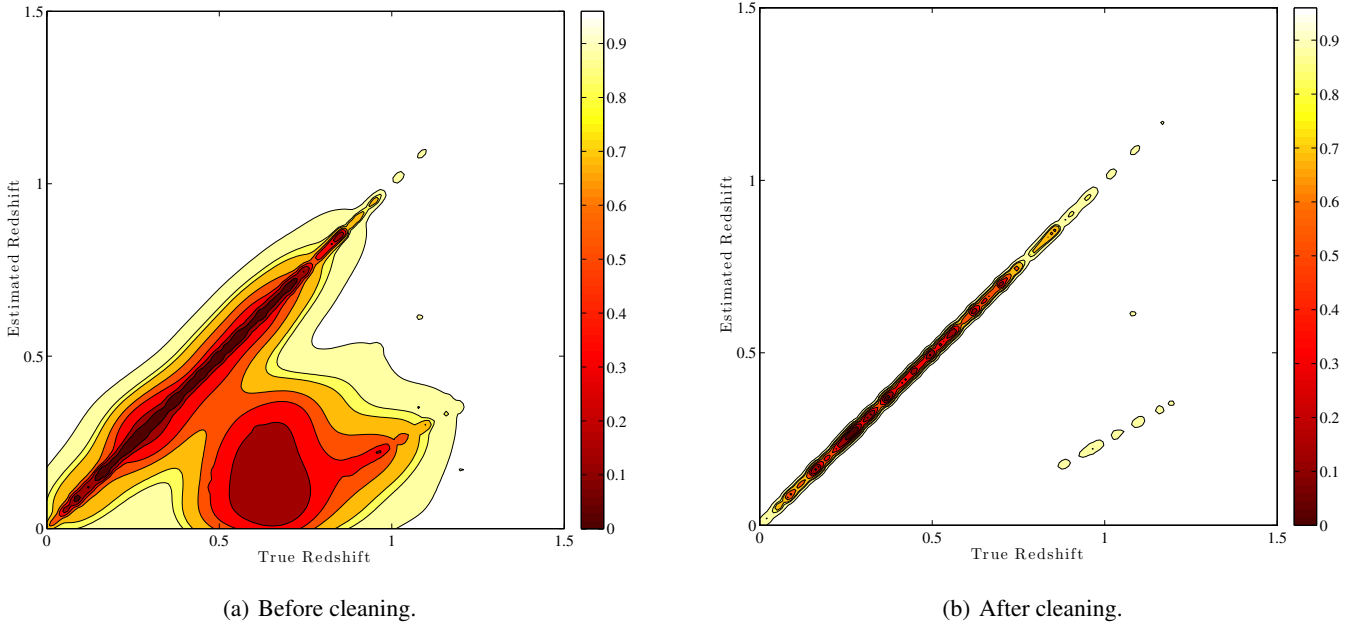
This represents a snapshot of how the Darth Fader algorithm works: we can *blindly* isolate a correct subset of galaxies, ensuring a good coverage of the correct data available in the pre-cleaned catalogue, and we can – by choosing an appropriate FDR threshold – guarantee that the resultant catalogue contains a very low catastrophic failure rate. Though not implemented in Darth Fader, the data rejected could be subject to further analysis, using additional information (e.g. photometry) and alternative methodology to determine the redshifts of the galaxies.

Figure 8 shows the catastrophic failure rate of Darth Fader before and after catalogue cleaning for a fixed FDR threshold of  $\alpha = 4.55\%$ , as a function of median S/N in the  $r$ -band. At high S/N ( $\sim 20$ ), the standard cross-correlation method yields a low catastrophic failure rate, and cleaning yields little improvement. At an S/N of 10, however, the standard cross-correlation method experiences a progressive increase in the catastrophic failure rate, approaching 50% at an S/N of 1; in contrast our method can maintain a low catastrophic failure rate ( $\lesssim 5\%$ ) for S/N levels of  $\geq 1$ .

An important point to note is that the catastrophic failure rate before cleaning ( $\mathcal{F}$ ) represents a theoretical minimum amount of data that *must* be discarded with a perfect catalogue cleaning method (where  $\mathcal{F}_c$  and  $C$  would be 0 and 100% respectively); thus the theoretical maximum retention is given by  $100\% - \mathcal{F}$ . In practice the amount discarded is usually greater (since we inevitably discard galaxies that would otherwise yield correct redshifts), but it can also be less than this if a more relaxed threshold is used, necessarily incurring false positive contamination in the retained data set.

Using an FDR threshold of 4.55% allowed false detections, a catalogue of  $S/N = 2$  has a catastrophic failure rate of 34.5% before cleaning, thus our maximum expected retention in a perfect catalogue cleaning should only be 65.5%, with our actual retention at that FDR threshold being 52.6%. It should therefore not be surprising that at the lower end of signal-to-noise the expected retention values for a cleaned catalogue are (necessarily) low; however this can still represent a large proportion of the *correct* data available. The recovery of these data still represents a net gain when compared to a total loss of these data, particularly when these recovered data can be known to be highly accurate.

At higher S/N, the impact of cleaning is reduced because denoising does not reveal more, significantly useful, diagnostic information: the number of features present in the noisy spectrum will more frequently already meet the selection criterion before cleaning, and thus cleaning the catalogue removes fewer spectra. To ensure a similarly low catastrophic failure rate in low S/N data would require a stricter FDR threshold to be used, and therefore would result in more data being discarded.



**Fig. 7.** A contour plot to show the effect on redshift estimation **a)** before and **b)** after cleaning a catalogue which is at a signal-to-noise of 2.0, and cleaned with an FDR threshold of 4.55% allowed false detections. Contours indicate the fraction of points enclosed within them. Just under two thirds of all the estimated redshifts lie on the diagonal (and are thus correct) before cleaning being applied. Clearly outliers still exist after cleaning of the catalogue (off-diagonal), where the redshift estimation has failed, but as it can be seen, these are very few, and the result has a high certainty, with 94.9% of the estimates being correct. The capture rate for this catalogue and at this FDR threshold is 76.2%.

To demonstrate the effect of the choice of FDR threshold on the catastrophic failure rate, the retention and the capture rate in the very low S/N regime, we test Darth Fader on two fixed low S/N catalogues of median S/N values of 1.0 and 2.0 with flat white-Gaussian noise, and one mixed S/N catalogue consisting of spectra with pixel-dependent Gaussian noise (see Sect. 6.1) with a uniform distribution of median S/N values between 1 and 20.

Figure 9 clearly demonstrates the tradeoff that exists between the catastrophic failure rate after cleaning and the capture rate. Relaxing the threshold (i.e. increasing  $\alpha$ ) improves both the retention and the capture rate by detecting more features in the spectra, more of which are likely to be false features rather than true ones, and thereby increasing the number of spectra accepted under the feature-counting criterion, but at a cost to the catastrophic failure rate since more erroneous spectra will also be accepted. A more conservative approach leads the FDR denoising to remove more real features, with the guarantee that very few of the remaining features will be false detections. This leads to a general decrease in both the retention and the capture rate since fewer spectra will exhibit the required number of features after denoising, with the benefit of this being a decrease in the catastrophic failure rate.

Notice also in Fig. 9 that beyond a certain point the catastrophic failure rate saturates for the spectra with white-Gaussian noise (and shows little improvement for the mixed S/N catalogue with pixel-dependent noise), and stricter FDR thresholding (resulting in a smaller fraction of false detections) does not yield significant reductions in the rate of catastrophic failures; indeed this only serves to penalise both retention and the capture rate.

The results of the uniformly mixed and pixel-dependent S/N catalogue represent a step toward a more realistic view of what a real galaxy survey could look like. A real survey would not, however, have such a uniform distribution of S/N values, and would be skewed toward a greater population at lower S/N, with the

actual distribution in signal-to-noise being specific to the type of survey and the instrument used.

## 6. Application to data

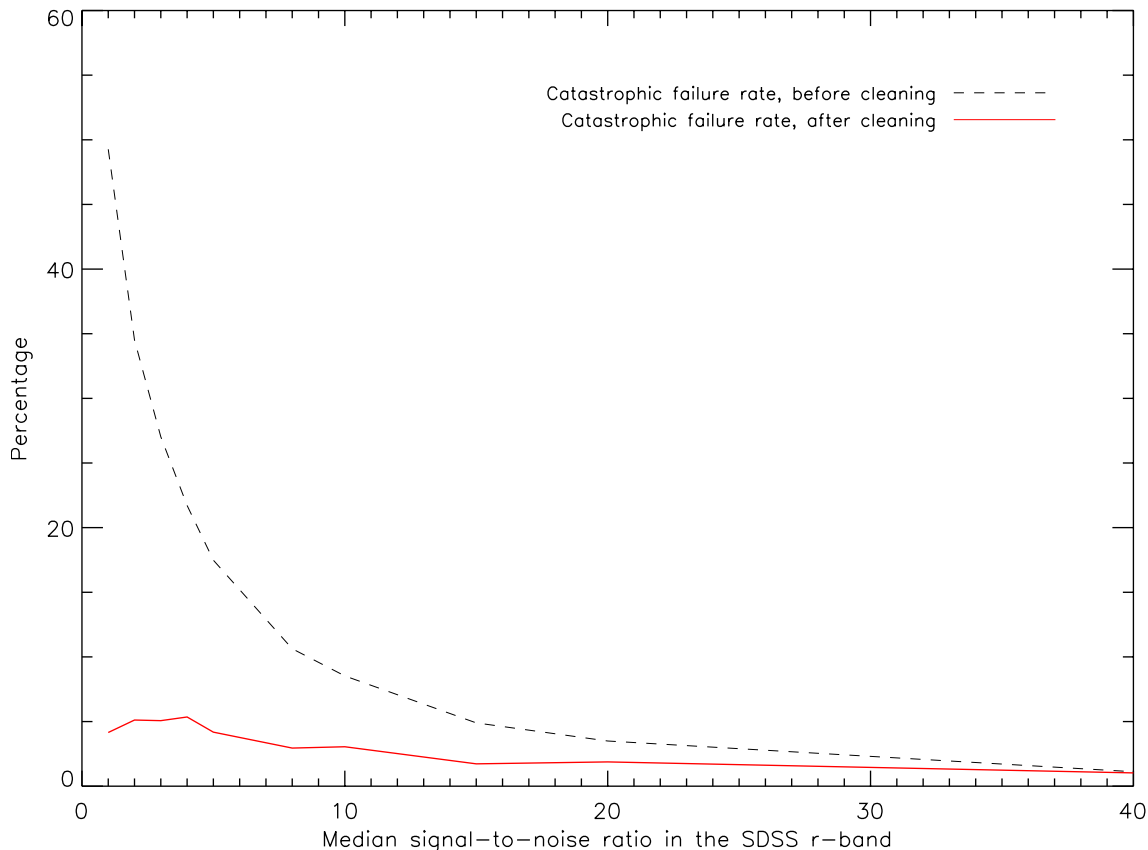
In the previous section, we demonstrated the robustness of the Darth Fader algorithm on simulations. Here we expand on the work in Sect. 5 to show that our feature detection methods work well on real spectra from the SDSS archive.

A full test on real SDSS data, for example, would not be practical since low S/N spectra are automatically discarded by magnitude cuts in the SDSS data processing pipeline and not available in their data products. Spectroscopic cuts for the main galaxy sample are taken at magnitudes in  $r < 17.77$  (Petrosian) and the resulting spectra have a median S/N (per pixel)  $> 4$  (Strauss et al. 2002).

Real data differ from our simulations in a number of important ways: rare galaxy types/properties may exist within real data catalogues, and these may not necessarily be well encompassed by our simulations, and real data can often have more complex noise properties. It is therefore important to test whether our denoising methods, and feature-counting criterion, can be applied to real data.

### 6.1. Realistic pixel-dependent noise

Real spectrographs have a sensitivity that varies – sometimes quite strongly – with wavelength or per pixel, primarily as a result of the sky brightness and instrumental efficiency. We simulate a realistic error-curve that spans the typical optical survey wavelength range, and in Fig. 10 we show the  $1\sigma$  error-curve per pixel used to mimic a realistic instrument. This is similar to what could be expected for an existing survey such as SDSS, or the forthcoming DESI spectroscopic survey



**Fig. 8.** Illustration of how Darth Fader improves the catastrophic failure rate of the redshift estimates of the test catalogue at different signal-to-noise values (flat white-Gaussian noise) for a fixed FDR threshold of 4.55% allowed false detections. Note the marked improvement in the S/N range 1.0–10.0 where catastrophic failure rates are reduced by up to 40%. For this choice of  $\alpha$ , the catastrophic failure rate is always found to be  $\leq 5\%$  after cleaning, for S/N values  $\geq 1$ . Our catastrophic failure rate after cleaning at an S/N of 1 is similar to the rate for an S/N value of 15 without cleaning. The catastrophic failure rate before cleaning (dashed line) represents the theoretical minimum amount of data that must be discarded for perfect catalogue cleaning.

(Levi et al. 2013), (itself a merger of the BigBOSS (Schlegel et al. 2011) & DESpec (Abdalla et al. 2012) spectroscopic surveys), as well as other projects involving multi-object spectrographs<sup>6</sup>. The Darth Fader algorithm can use the error-curve in the denoising step. Better accounting for the complex noise properties of the observed spectrum enhances the ability to discriminate between true features and those arising due to noise.

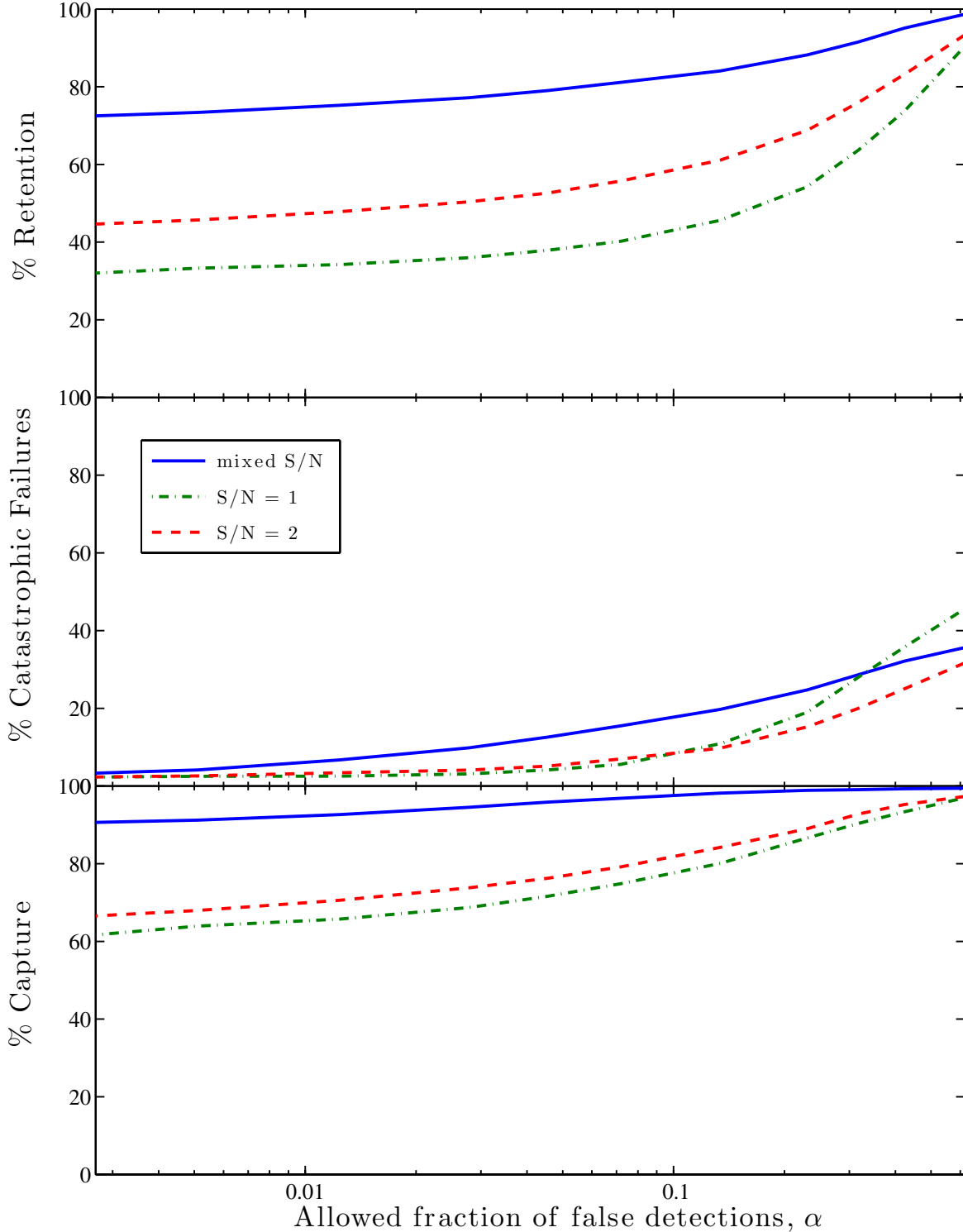
Figure 11 shows a continuum-subtracted spectrum from our catalogue, truncated to match the wavelength range of the error-curve (and hence our simulated instrument), before and after the addition of the wavelength-dependent noise of Fig. 10. We also plot the spectrum after denoising (again with an FDR threshold of 4.55% allowed false detections) with the Darth Fader algorithm when supplied with the error-curve in Fig. 10. This spectrum has a median S/N of 5 in the  $r$ -band at this particular redshift, ( $z = 1.5$ ). However, for the same noise level, this S/N would vary between 3 and 5 according to redshift, as a result of different continuum levels within the boundaries of the  $r$ -band.

To test the effectiveness and robustness of the denoising, we use the same test spectrum as in Fig. 11, and apply 10 000 random (wrap-around) shifts in order to randomise the location of the principal features. For each shifted spectrum, pixel-dependent Gaussian noise is added as before, and at the same

level. We then perform a denoising on each spectrum, and compute the residual with the input noisy spectrum. The RMS residual gives an estimate of the noise with its statistical distribution – if the denoising has been effective – matching the input error-curve. The randomised shifting of the spectrum allows us to determine the effectiveness of the denoising independently of the locations of the true features, and removes any cumulative biasing arising from features being undetected after denoising, or denoising artefacts. We do however expect a biasing at the edges of the spectrum at both the long and short-wavelength ends, due to a lack of information “beyond” the edge limiting the ability to correctly characterise local features at the edge as either signal or noise. In Fig. 12, we show the ratio of the noise standard deviation over the input error-curve as a function of wavelength, for both the pixel-dependent noise in Fig. 10, at FDR parameters of  $\alpha = 4.55\%$  and  $\alpha = 0.27\%$ , and for flat white Gaussian noise ( $\alpha = 4.55\%$ ). The noise standard deviation has been computed from the 10, 000 residuals described above.

As can be seen in Fig. 12, the addition and subsequent denoising of flat noise behaves as one would expect; small deviations about a flat line situated at  $y = 1$  (shifted down in the figure for clarity). Minor artefacts are present at the edges due to border effects specific to the wavelet transform, and features occasionally straddling the edges of the spectrum. The more complicated noise proves to be a considerably more difficult task than the flat noise, and clearly has some persistent residual

<sup>6</sup> These surveys are, however, expected to be at much higher resolution than our simulations.

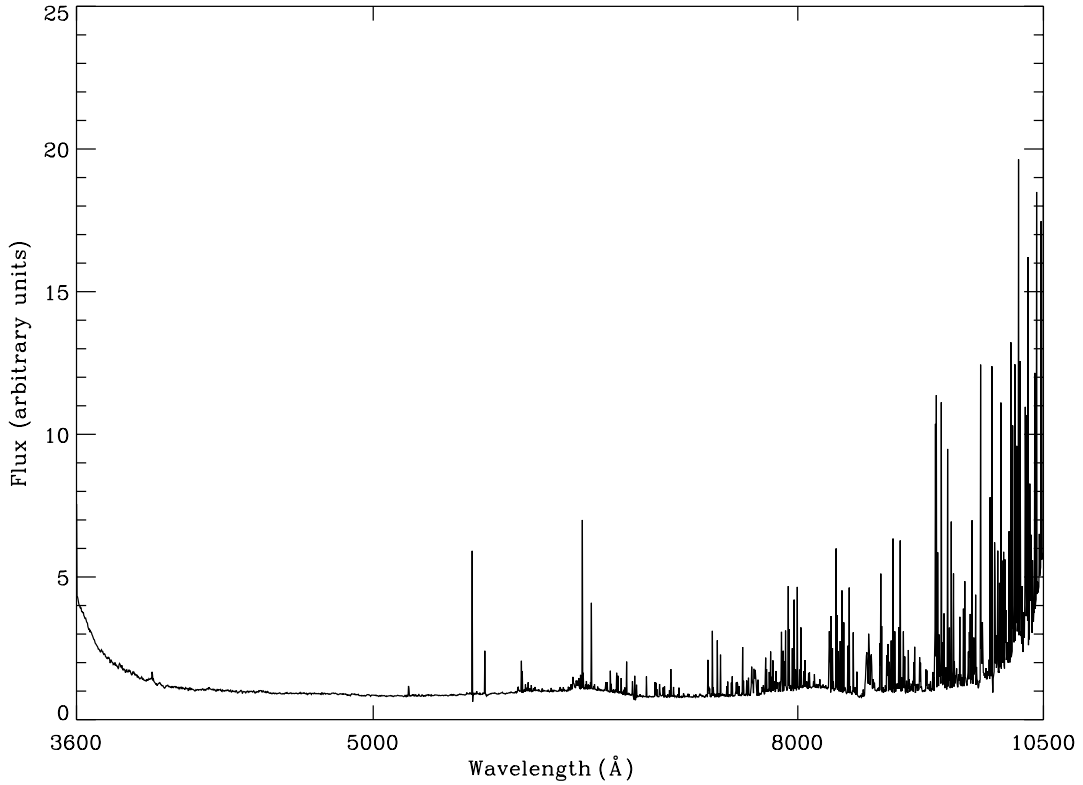


**Fig. 9.** Effect of the choice of FDR threshold on the catastrophic failure rate after cleaning (*upper panel*), the retention (*middle panel*) and the capture rate (*lower panel*) on catalogues with a fixed S/N of 1.0 and 2.0 with flat noise; and on a mixed catalogue with a uniform distribution in S/N between 1 and 20, with pixel-dependent noise. Note the greater sacrifices required both in retention and capture rate in order to obtain the same catastrophic failure rate at an S/N of 1.0 compared to 2.0. Note also that we are able to obtain a 5.1% failure rate in our redshift estimates for the cleaned catalogue, a retention of 52.6%, and a capture rate of 76.2% with the catalogue at an S/N of 2 at an FDR threshold of 4.55%.

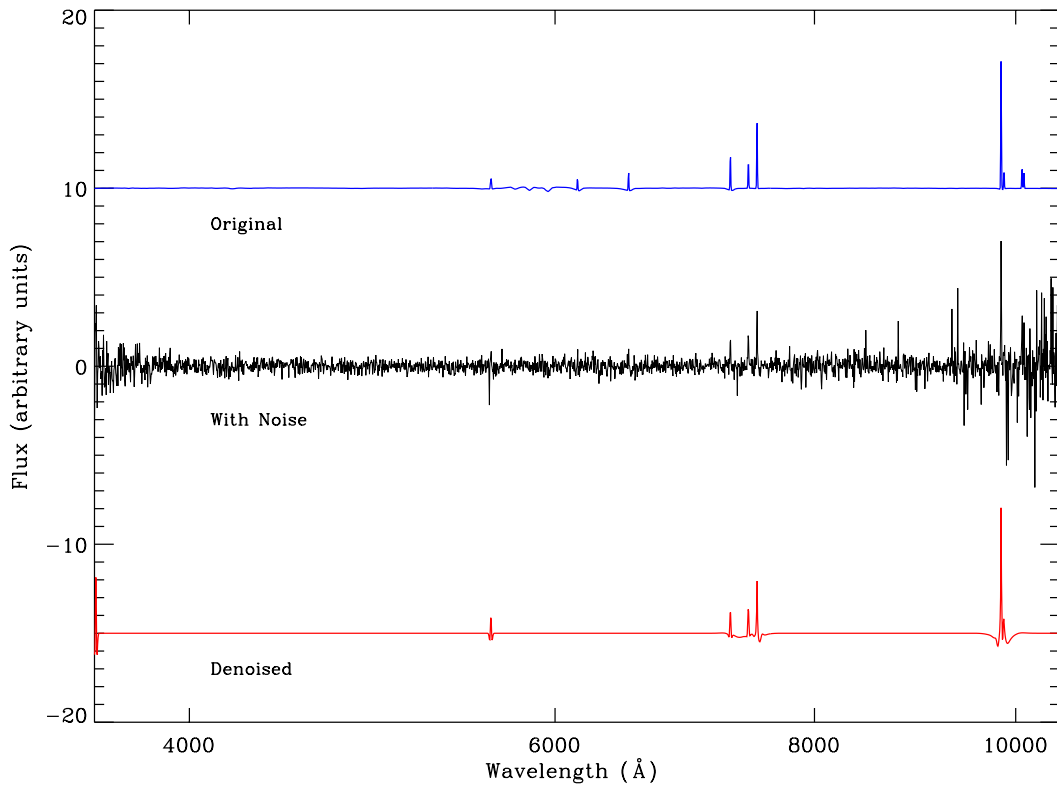
features after denoising, particularly in the longer wavelength range where the error-curve is most complex. This discrepancy is due to the denoising not fully accounting for the rapidly changing noise from one pixel to the next.

Clearly this will impact on feature detection, resulting in a greater number spurious detections particularly at longer

wavelengths. Increasing the FDR parameter,  $\alpha$ , does provide significant improvement in the efficacy of the denoising (as shown by the middle curve). It may be possible to further ameliorate these systematic effects with a more optimal wavelet choice (should one exist), or by assigning a weight to each pixel to counterbalance the effect of the denoising not fully accounting

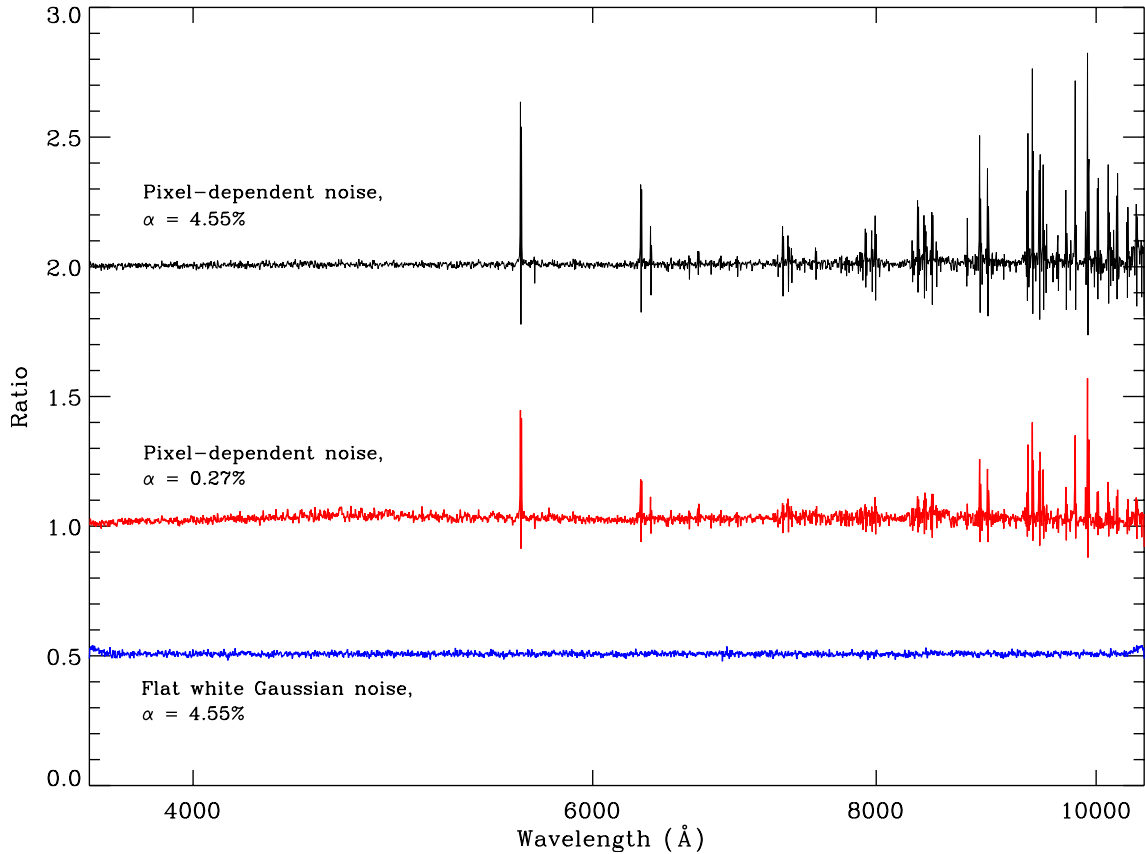


**Fig. 10.** A realistic error-curve, where the resolution and binning are the same as for our mock catalogue, but with the wavelength range being slightly shorter, in order to be more proximal to the wavelength range of a realistic instrument. Gaussian noise is added to each pixel in our simulated data, with a standard deviation given by the value of the error-curve at that same pixel.



**Fig. 11.** Denoising of test spectrum (cf. Fig. 2, continuum-subtracted) with pixel-dependent noise. Note how most of the main features are detected and how, for this particular noise realisation, no false detections are found in the complicated and noisy long-wavelength region. We do incur a false detection at the very short-wavelength end of the spectrum. This is a systematic edge-effect resulting from a lack of information that would otherwise allow the algorithm to properly distinguish this as a noise feature.





**Fig. 12.** Ratio of the true error-curve with respect to the derived error-curve from the rms error per pixel on the difference between the original input spectrum and the denoised spectrum for both flat noise and pixel-dependent noise. The lower curve (blue) has been shifted down (by 0.5) for clarity, and the upper curve (black), has also been shifted up (by 1.0) for clarity. Note the minor systematic edge effects on the denoising of white-Gaussian (flat) noise. Clearly the complex noise region has a marked systematic effect on the denoising, with rapidly changing noise regions experiencing both over- and under-estimates in the noise strength. This systematic effect is dependent upon the FDR threshold chosen, with thresholding that is less strict (upper curve) being more prone than stricter thresholding (middle curve).

for the complex noise properties. Additionally, Fig. 9 (solid blue line) shows that a stricter FDR thresholding is already effective in mitigating these systematic effects of the denoising.

We can conclude therefore that DARTH FADER provides effective and robust denoising, regardless of whether the noise is stationary or wavelength dependent, provided that a choice of FDR parameter is made such that it is appropriate to the type of noise present.

## 6.2. Feature extraction in real data

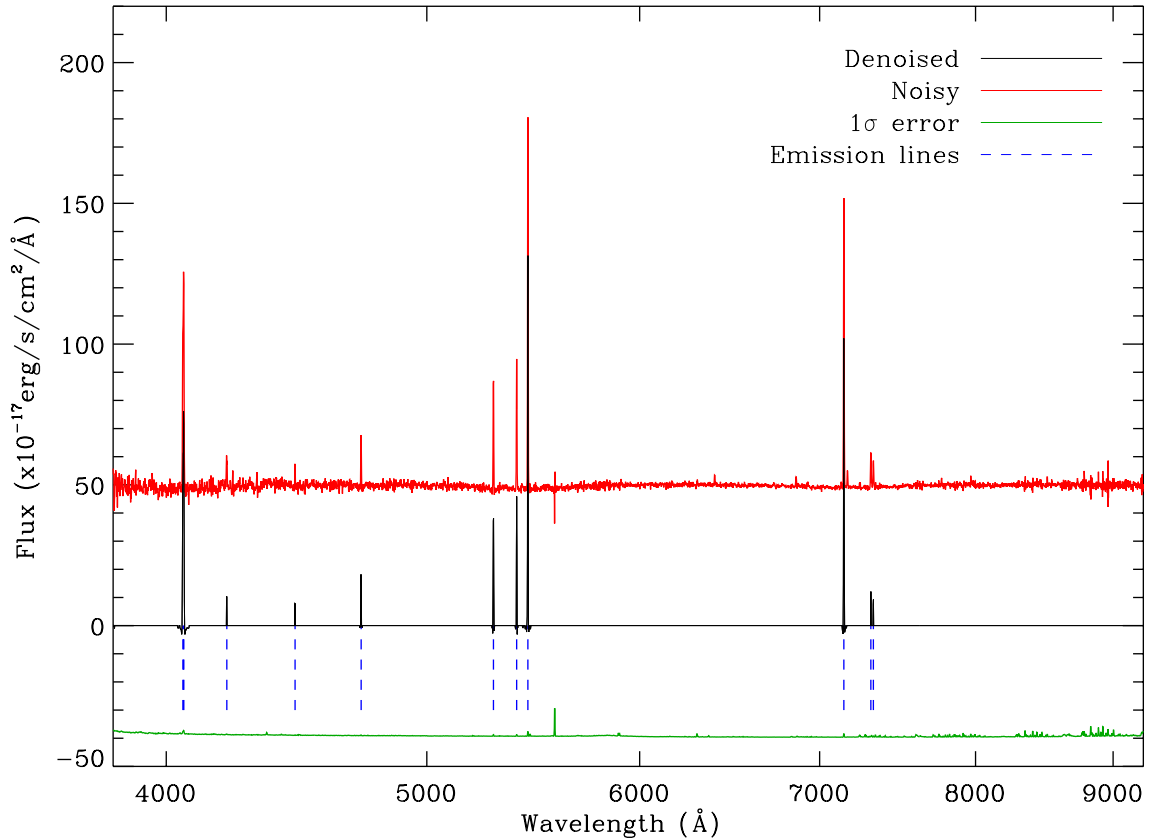
The number of features we can detect stems in part from the resolution of the spectra: if the resolution is poor, there is more uncertainty in the precise wavelength of the spectral lines, making it easier to confuse lines because they are slightly smeared out. Another, more important concern, is the potential blending of doublets or other lines in close proximity in wavelength, such as [N II] with  $H_{\alpha}$ . These localised groupings of lines provide powerful constraints on the galaxy redshift since the wavelength gap between the two lines in a doublet or close pair is often sufficient to conclusively isolate which emission lines are present, and hence deduce the redshift. Poorer resolution will often result in blending of such features, limiting the number of detectable features as well as broadening the possible location of the feature. Hence poor resolution impacts both the number of features

through blending, and the detected locations of the features in wavelength due to coarser pixelisation of the data.

This can reduce the number of spectra meeting the feature-counting criterion in poorer resolution spectra, however this might be mitigated by considering a larger wavelength range for the spectra: provided features exist in this extended range, more features can be found to counteract the loss of feature detections and precision as a consequence of the poorer resolution. It is for this reason that our simulated spectra cover a larger wavelength range than SDSS currently does (however DESI is expected to have a similar wavelength range). In reality this trade-off is a minor consideration, however, since the practicalities of instrument design are the limiting factor for the wavelength range of spectra in real surveys. Our simulated spectra are at moderate resolution; instruments such as SDSS offer substantially higher resolution spectra.

In order to show the broader applicability of DARTH FADER to real spectra potentially at higher resolution and covering a narrower wavelength range, we take three SDSS galaxy spectra (an emission line galaxy, ELG, a luminous red galaxy, LRG, and a “typical” galaxy) and their respective  $1\sigma$  error-curves, as fiducial type galaxies that well represent the SDSS galaxy catalogue<sup>7</sup>.

<sup>7</sup> These three example galaxies can be found at: [http://www.sdss.org/gallery/gal\\_spectra.html](http://www.sdss.org/gallery/gal_spectra.html) with Plate ID: 312, MJD: 51689 and Fibre IDs: 220 (LRG), 255 (Typical), & 529 (ELG).



**Fig. 13.** Denoising and feature extraction for an SDSS ELG. The noisy spectrum (red) has been shifted up, and the error-curve (green) shifted down, for clarity. The vertical dashed lines (blue) indicate the locations of detected features that correspond to true emission features. The FDR denoising and feature extraction clearly pinpoints all of the major features without any difficulty. The three largest lines are, from left to right, the [O II] doublet, [O III] and  $H_{\alpha}$ .

These spectra have a resolution  $R(=\lambda/\Delta\lambda)$  significantly higher than that of our simulations, namely  $R \sim 1845$  compared to  $R \sim 850$ , and as such features are better separated. The  $r$ -band S/N for these galaxies is quoted to be 9.2 for the ELG, 9.3 for the LRG, and 15.0 for the typical galaxy, respectively.

In denoising these spectra we use an FDR threshold of  $\alpha = 0.27\%$  as motivated by the discussion in Sect. 6.1 and the results in Fig. 9. We apply a positivity (and “negativity”) constraint, as before, to denoise the positive and negative sections of each spectrum independently, and recombine them to form the final denoised spectrum. The procedure uses the same positivity constraint, once on denoising the spectrum, and once on denoising the reverse-signed spectrum – this is entirely equivalent to denoising once with a positivity constraint, and again with a “negativity” constraint.

In Fig. 13, we show the continuum-subtracted spectrum, the FDR denoised spectrum, and the line features we detect for the emission-line galaxy. We also plot the  $1\sigma$  error-curve, which we assume as Gaussian.

This ELG spectrum has many strong features, so it is not surprising that the FDR denoising detects most of them. We do however miss one very weak emission feature that is comparable to the noise, at  $\sim 7800 \text{ \AA}$ . It should also be noted that the potential line-like features arising from the noise, namely at  $\sim 5600 \text{ \AA}$  and again at  $\sim 8950 \text{ \AA}$  are completely ignored by the FDR denoising since, by supplying the error-curve, these features are correctly identified as likely arising from noise rather than signal.

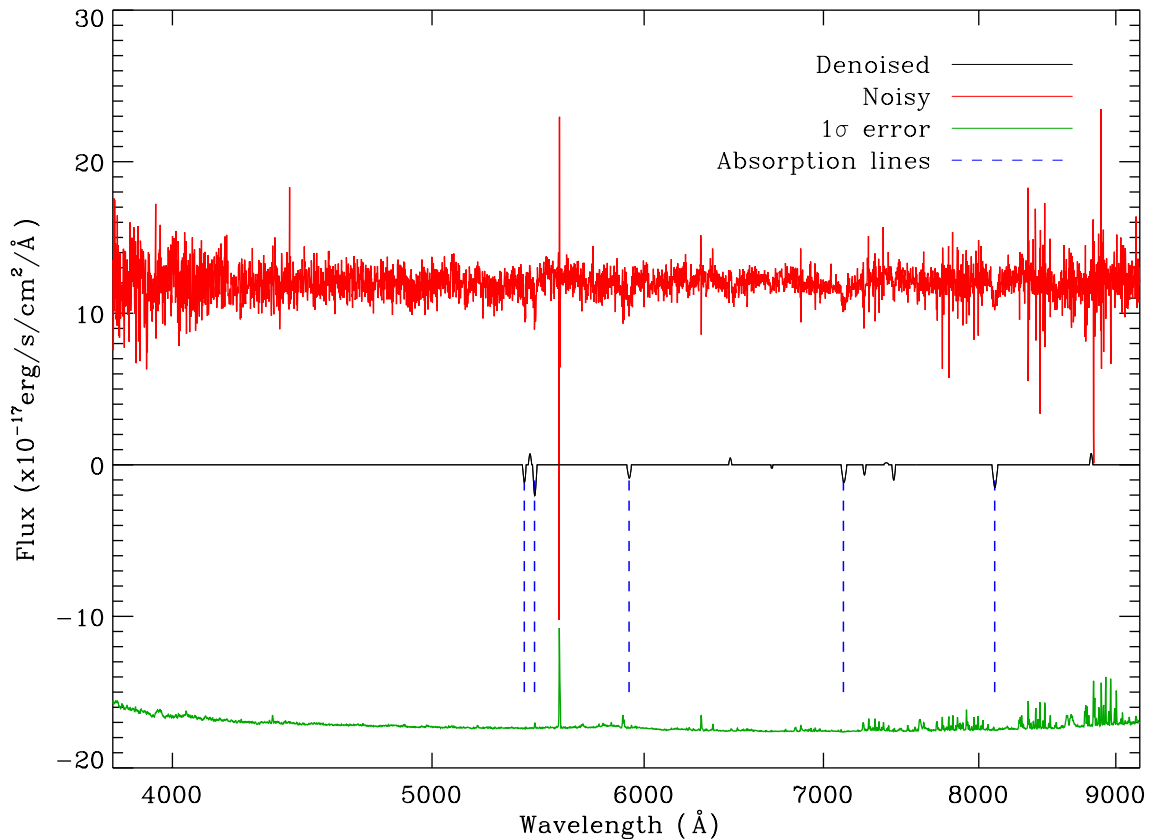
Feature detection in the LRG spectrum (Fig. 14) presents a more difficult challenge. Despite the signal-to-noise ratio in the  $r$ -band being of similar value to the ELG, the widths of the features compared to the noise (i.e. the signal-to-noise values on the lines) are much smaller. We successfully detect five absorption features, despite them not being particularly prominent. We detect further spurious, smaller, features that cannot be associated with any common lines, and are likely minor artefacts from denoising.

The results for the typical galaxy are similar to those of the LRG (Fig. 15). In this case, we again detect all five of the absorption features, in addition we obtain some unidentifiable spurious features.

For each of the galaxy types shown here we can detect at least six features, though not all of them are true detections, nor do we require them to be necessarily true detections. Though we only consider Gaussian noise here, the tools used in Darth Fader are in principle not limited to purely Gaussian errors, and can be utilised with different types of errors (in particular Poisson, Gaussian + Poisson, multiplicative and correlated errors), and provided that denoising can be done appropriately the impact on the Darth Fader method will be minimal.

## 7. 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.



**Fig. 14.** Denoising and feature extraction for an SDSS LRG. The absorption lines from left to right are CaII (H and K), *G*-band, MgI and NaI. *G*-band absorption is not strictly an absorption *line*, but rather an aggregate absorption feature due to the presence of multiple lines arising from metals (mainly iron) in the numerous G-type stars present in the galaxy population. Also not to be confused with the SDSS photometric filter *g*-band.

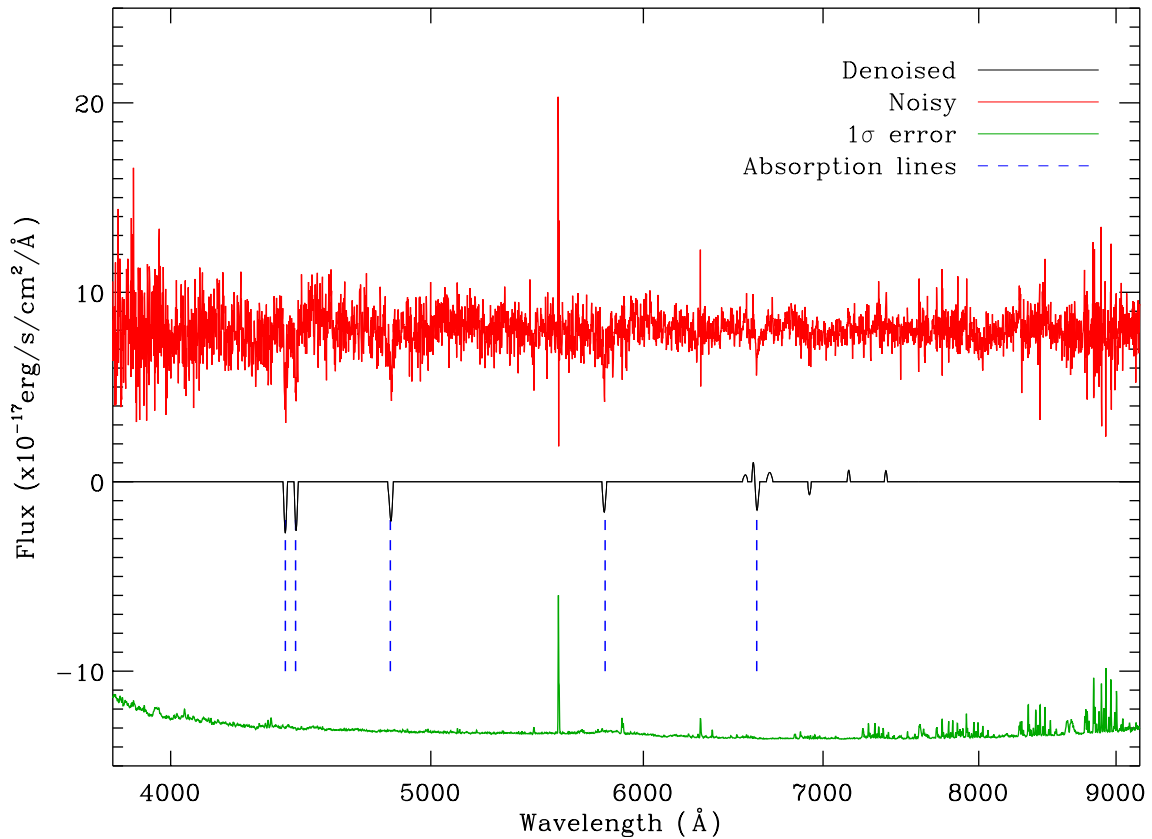
We can successfully make an estimate of the continuum without needing to model the spectra, and we can confidently make use of data at signal-to-noise levels that were previously beyond the reach of other techniques. This is achieved by denoising the data with an appropriately chosen false detection rate threshold and implementing a simple feature-counting criterion, resulting in very low catastrophic failure rate for 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 extract what is likely to be good data for redshift estimation from what is likely to yield an inaccurate redshift estimate, with a good level of confidence. Even at signal-to-noise levels as low as 2.0 in the *r*-band, we can retain 52.6% of the data, and contained within this subset we can obtain 76.2% of all the potentially correct redshift estimates that were initially available, resulting in a highly confident subsample where 94.9% of the redshift estimates are reliable. This cleaning therefore has applications in large surveys, such as the upcoming Euclid survey (Refregier et al. 2010; Laureijs et al. 2011), which requires a spectroscopic redshift catalogue with very few catastrophic failures.

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 tend to be those at higher redshifts. Currently, lower signal to noise spectra tend to be discarded, or to yield highly unreliable redshift estimates. DARTH FADER allows

for the inclusion of a substantial subset of these otherwise-discarded galaxies.

DARTH FADER demonstrates that these current methods, with a blunt cut-off in the signal-to-noise/flux for what is considered to be informative data, can be significantly improved upon, and that improvements are available all the way down to very low signal-to-noise levels. The levels of retention presented in this paper may seem moderate, however, for such low signal-to-noise data they can only be expected to be so, as redshift estimation necessarily fails for a large fraction of spectra at these high noise levels. DARTH FADER performs very well and can reliably extract the majority of the likely correct data that is available in the low S/N regime. It is for this reason that for our method the capture rate is a more useful diagnostic than the retention, since it shows that for the available informative data contained within the uninformative data, DARTH FADER can isolate and extract the majority of it in a blind and fully automated manner, resulting in catalogues where we can have a very high degree of certainty that the redshifts are correct. Indeed, for all the S/N levels we test, DARTH FADER is always able to capture at least 60% of the available data that we know to be correct, with this minimum rising to 70% with more standard choices of FDR parameter values (~4%). Hence – even when overall retention values appear moderate – the low rate of catastrophic failures together with the high proportion of good data available that is retained, represent a substantial gain when the alternative is to throw away the entire dataset with a blunt S/N cut because we cannot be certain as to which spectra are viable or not.



**Fig. 15.** Denoising and feature extraction for an SDSS typical galaxy. This spectrum is similar to that of the LRG, the highlighted absorption lines being the same as previously.

In addition, Darth Fader can deal with realistic noise providing we know the  $1\sigma$  error-curve, where we can reduce catastrophic failure rates of a mixed S/N catalogue from 22.7% down to 3.3% with a capture rate of 90.6% with an appropriate choice of FDR parameter. The algorithm does a good job of capturing nearly all of the spectra for which we are able to obtain a correct redshift estimate, and simultaneously maintaining a low catastrophic failure rate. Furthermore, we have shown that the continuum subtraction and feature-identification methods used in Darth Fader are effective on spectra from the SDSS data archive. This provides a proof-of-concept of the applicability of the method to real data, though there may be scope for development in this area. An additional feature is that, depending on different needs, the FDR parameter can be adjusted to enhance the capture rate and retention at a small cost to the catastrophic failure rate, or vice versa.

Our analysis in this paper utilises catalogues which are simulated, and as such they may be simpler than catalogues of real data; they do however fully represent the expected various galactic morphological types, distribution and redshift properties. The noise properties we use in our simulations may also be less complicated than those in real data, however, non-stationary Gaussian noise (varying per pixel) is a good enough approximation to real data, and this has been shown with the competent denoising of the SDSS example spectra (Figs. 13–15).

The simulations we use are of considerably lower resolution ( $R \sim 850$  compared to  $R \sim 1845$ ) than would be expected for a modern day spectral survey, with SDSS resolution being over twice as high, and the forthcoming DESI survey (Levi et al. 2013) – a merger of the BigBOSS and DESpec

surveys (Schlegel et al. 2011; Abdalla et al. 2012, respectively) – expected to be higher still. The wavelength range of our simulated spectra (3000 Å to 10 500 Å) are slightly longer than would be expected for a realistic instrument ( $3500 \lesssim \lambda \lesssim 10\,000$ ), but given the poorer resolution in our simulations, it is justifiable to extend the range. These factors do not, however, prevent these catalogues from being realistic. Indeed we have shown that it is possible to detect the required number of features in a shorter wavelength range, such as that of SDSS. There is therefore great promise for the use of these techniques in future large scale structure surveys, for feature extraction, redshift determination, and photometric calibration.

Continuum removal with wavelets, when compared against elaborate modelling, may be seen as a comparatively naïve 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.

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. Indeed, for any spectra whose components are easily modelled with the correct choice of wavelet, we expect our continuum subtraction method to work as demonstrated.

Although we only consider the numbers of features in this paper, the ability of the Darth Fader algorithm to detect likely true lines could readily be adapted to deal with feature *identification*, in particular for spectra where noise levels are very high, and  $K\sigma$  clipping would offer little advantage (since it has

a tendency to clip signal as well as noise). This would make it possible to cross-check the position of standard lines which have been redshifted to match the estimated redshift of the galaxy, against the positions of the maxima in the FDR denoised spectrum (since these are the features considered important by FDR).

Darth Fader is clearly useful for both redshift estimation and empirical continuum estimation and will be made publicly available as part of the iSAP<sup>8</sup> suite of codes. The blind nature of our algorithm, together with the ability to handle realistic noise, show promise for its inclusion in future spectral survey pipelines and data analyses.

*Acknowledgements.* The authors would like to acknowledge the support provided by the European Research Council, through grant SparseAstro (ERC-228261), F.B.A. acknowledges support from the Royal Society via a URF. We also wish to thank J  r  my Rapin & Simon Beckouche for their helpful feedback and assistance in the data presentation. The authors would also like to thank the anonymous referee for his/her thoughts and contributions, which have helped to improve the quality of this paper.

## References

- Abdalla, F., Annis, J., Bacon, D., et al. 2012 [[arXiv:1209.2451](https://arxiv.org/abs/1209.2451)]  
 Aihara, H., Allende Prieto, C., An, D., et al. 2011, *ApJS*, 193, 29  
 Benjamini, Y., & Hochberg, Y. 1995, *J. R. Stat. Soc. B*, 57, 289  
 Bolton, A. S., Schlegel, D. J., Aubourg,   ., et al. 2012, *AJ*, 144, 144  
 Bruzual, G., & Charlot, S. 2003, *MNRAS*, 344, 1000  
 Capak, P. L. 2009, in *AAS Meeting Abstracts*, #214, #200.06, *BAAS*, 41, 708  
 Costero, R., & Osterbrock, D. E. 1977, *ApJ*, 211, 675  
 Fadili, M. J., & Starck, J.-L. 2009, *IEEE ICIP 2009*, 1461  
 Fligge, M., & Solanki, S. K. 1997, *A&AS*, 124, 579  
 Fukugita, M., Ichikawa, T., Gunn, J. E., et al. 1996, *AJ*, 111, 1748  
 Garilli, B., Fumana, M., Franzetti, P., et al. 2010, *PASP*, 122, 827  
 Glazebrook, K., Offer, A. R., & Deeley, K. 1998, *ApJ*, 492, 98  
 Hopkins, A. M., Miller, C. J., Connolly, A. J., et al. 2002, *ApJ*, 123, 1086  
 Ilbert, O., Capak, P., Salvato, M., et al. 2009, *ApJ*, 690, 1236  
 Jouvel, S., Kneib, J.-P., Ilbert, O., et al. 2009, *A&A*, 504, 359  
 Koski, A. T., & Osterbrock, D. E. 1976, *ApJ*, 203, L49  
 Kurtz, M. J., & Mink, D. J. 1998, *PASP*, 110, 934  
 Laureijs, R., Amiaux, J., Arduini, S., et al. 2011 [[arXiv:1110.3193](https://arxiv.org/abs/1110.3193)]  
 Le F  vre, O., Crampton, D., Lilly, S. J., Hammer, F., & Tresse, L. 1995, *ApJ*, 455, 60  
 Le F  vre, O., Vettolani, G., Garilli, B., et al. 2005, *A&A*, 439, 845  
 Levi, M., Bebek, C., Beers, T., et al. 2013 [[arXiv:1308.0847](https://arxiv.org/abs/1308.0847)]  
 Lutz, R., Schuh, S., Silvotti, R., et al. 2008, in *Hot Subdwarf Stars and Related Objects*, eds. U. Heber, C. S. Jeffery, & R. Napiwotzki, *ASP Conf. Ser.*, 392, 339  
 Miller, C. J., Genovese, C., Nichol, R. C., et al. 2001, *ApJ*, 122, 3492  
 Panuzzo, P., Vega, O., Bressan, A., et al. 2007, *ApJ*, 656, 206  
 Pires, S., Juin, J. B., Yvon, D., et al. 2006, *A&A*, 455, 741  
 Refregier, A., Amara, A., Kitching, T. D., et al. 2010 [[arXiv:1001.0061](https://arxiv.org/abs/1001.0061)]  
 Schlegel, D., Abdalla, F., Abraham, T., et al. 2011 [[arXiv:1106.1706](https://arxiv.org/abs/1106.1706)]  
 Smee, S., Gunn, J. E., Uomoto, A., et al. 2013, *Ap*, 146, 32  
 Starck, J.-L., & Murtagh, F. 1994, *A&A*, 288, 342  
 Starck, J.-L., & Murtagh, F. 2006, *Astronomical Image and Data Analysis*, 2nd edn. (Springer)  
 Starck, J.-L., Bijaoui, A., & Murtagh, F. 1995, *CVGIP: Graphical Models and Image Processing*, 57, 420  
 Starck, J.-L., Claret, A., & Siebenmorgen, R. 1996a, *ISOCAM Data Calibration*, Tech. rep., CEA  
 Starck, J.-L., Murtagh, F., Pirenne, B., & Albrecht, M. 1996b, *PASP*, 108, 446  
 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, *AJ*, 123, 485  
 Strauss, M. A., Weinberg, D. H., Lupton, R. H., et al. 2002, *AJ*, 124, 1810  
 SubbaRao, M., Frieman, J., Bernardi, M., et al. 2002, *SPIE Proc.* 4847, eds. J.-L. Starck, & F. D. Murtagh, 452  
 Tonry, J., & Davis, M. 1979, *AJ*, 84, 1511  
 Yamada, I. 2001, in *Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications*, eds. D. Butnariu, Y. Censor, & S. Reich (Elsevier)

<sup>8</sup> iSAP package available at:  
<http://www.cosmostat.org/software.html>