



HAL
open science

Tracking the Lyman alpha emission line in the CircumGalactic Medium in MUSE data

Raphael Bacher, Pierre Maho, Florent Chatelain, Olivier J.J. Michel

► **To cite this version:**

Raphael Bacher, Pierre Maho, Florent Chatelain, Olivier J.J. Michel. Tracking the Lyman alpha emission line in the CircumGalactic Medium in MUSE data. D. Mary; R. Flamary; C. Theys and C. Aime. *Mathematical Tools for Instrumentation & Signal Processing in Astronomy*, 78-79, EDP Sciences, pp.233-245, 2016, EAS Publications Series, 978-2-7598-9006-4. 10.1051/eas/1678011 . hal-01503524

HAL Id: hal-01503524

<https://hal.science/hal-01503524>

Submitted on 18 Jul 2017

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.

TRACKING THE LYMAN ALPHA EMISSION LINE IN THE CIRCUMGALACTIC MEDIUM IN MUSE DATA

Raphael Bacher^{1,2}, Pierre Maho¹, Florent Chatelain¹ and Olivier Michel¹

Abstract. Since 2014, the Multi Unit Spectroscopic Explorer (MUSE) instrument generates hyperspectral datacubes (300 by 300 pixels by 3600 wavelength in the visible range) of the deep Universe. One of the main purposes of the wide field spectrograph MUSE is to analyse galaxies and their surroundings by the study of their spectra. Galaxy spectra are composed of a continuum emission and of sparse emission (or absorption) peaks. On the contrary surrounding gas only contains peak such as the Lyman alpha emission line. Several methods are developed here to detect the gas signature as far as possible in the galaxy surroundings. These methods combined clustering approaches and several pre-processing steps.

1 Introduction

One of the most abundant element of the Universe is hydrogen gas. It is believed that young galaxies are often surrounded by halos of hydrogen gas, the Circum Galactic Medium (CGM). The emissions from these halos can be several order of magnitude fainter than those from the galaxies. Furthermore, the emission spectrum of the former are narrow lines such as the Lyman- α line, illustrated in Figure 1.

In order to detect these faint halos, it is necessary to focus on narrow spectral bands around the emission line. However, in the presence of redshift, the spectral position of the Lyman- α in the measured spectrum cannot be known thus precluding to design an adapted narrow-band filter to detect faint Lyman- α emission *a priori*. Consequently, detecting these emissions requires to use a spectrograph that covers a wide band of the spectrum with a high resolution. Furthermore,

The authors would like to thanks the ERC grant 339659-MUSICOS for its funding

¹ GIPSA-Lab, UMR CNRS 5216 Grenoble University, 38400 Saint Martin d’Heres, France
email: raphael.bacher@gipsa-lab.fr

² CRAL, UMR CNRS 5574, 9 avenue Charles-Andre, 69230 Saint Genis-Laval, France

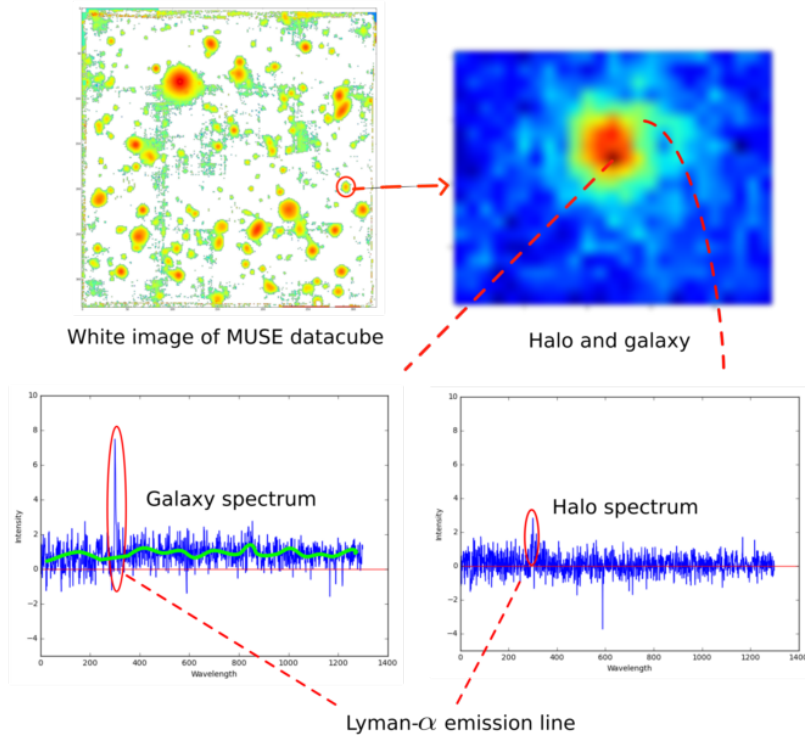


Fig. 1. Galaxy and halo spectra in MUSE data with Lyman- α peak and continuum emission.

exploring or analyzing the spatial neighborhood requests that the instrument have the largest possible field of view. The Multi-Unit Spectroscopic Explorer or MUSE (Bacon et al., 2010) was developed to respond to such needs : resulting of ten years of work of an European consortium led by the Centre de Recherches Astrophysique de Lyon (CRAL), it is a 3D spectrograph capable of imaging and analyzing a field of 1 arcmin^2 by producing an hyperspectral data cube of 300 by 300 pixels by 3600 wavelengths. Its spectral range covers the visible domain, from 450nm to 930nm with a spectral resolution $R = 2000 \dots 4000$.

Since MUSE first light in January 2014, several studies (Paris et al., 2013), (Meilrier et al., 2016) have already been conducted on MUSE data, with the aim to detect faint young galaxies. These latter are characterized by the presence a powerful Lyman- α emission line in their emission spectrum. The proposed detection methods mostly assume spatially and spectrally punctual sources; as a consequence, they efficiently find the core of galaxies but are not (yet) adapted to detect faint

extended halos. The purpose here is to explore the vicinity of these already detected galaxies, and track the Lyman- α emission line as far away from the galaxy as possible. Focusing the detection of Lyman- α emission line in both spectral and spatial neighborhood of an already detected (young) galaxy leads to a much smaller exploration area, making the problem somewhat "easier". However the researched signal can be several order of magnitude fainter than the known galaxy. The main issue for this problem can thus be reformulated as follows :

\Rightarrow Given a known galaxy with a continuum spectrum and a Lyman- α emission line, is there an pure Lyman- α emission spatially more extended than the galaxy spatial support ?

A first attempt to address the detection of halos in the Hubble Deep Field South (HDFS) data was recently published (Wisotzki et al., 2015). As the data recently produced by MUSE exhibit much higher spectral resolution (but lower spatial resolution), the halo detection problem can be revisited with alternate approaches, with the perspective to obtain improved performances.

Up to our knowledge, the study in (Courbot et al., 2016) is presently the only one addressing the problem of faint emission detection in hyperspectral data. The present paper addresses the same issues from a different perspective, relying on an unsupervised machine learning frame instead of a hypothesis test over a given dictionary. The goal here is to assert the presence of a halo without any *a priori* other than the similarity in the narrow spectral band between the galaxy center and the gas halo.

In section 2 both MUSE data and the Data Reduction System used to preprocess the data before its analysis are presented. The proposed clustering approach is described in section 3 and some results highlighting the performance of the proposed approach are given in section 4. Finally some other undergoing approaches are quickly exposed with some perspectives in section 5.

2 Data

MUSE produces huge amounts of data (a single night of observation can generate around 50 Go of raw data) that have to be processed by a Data Reduction System (DRS) before they can be used for science analysis. The DRS is briefly described below.

2.1 Acquisition

To be able to analyse the emitted light over a large field of view, MUSE is composed of 24 Integral Field Unit (IFU). Each IFU is dedicated to a part of the image, and creates 48 slices of data to be analyzed by a spectrograph. This process is summarized in Figure 2 and thoroughly described in (Weilbacher et al., 2012).

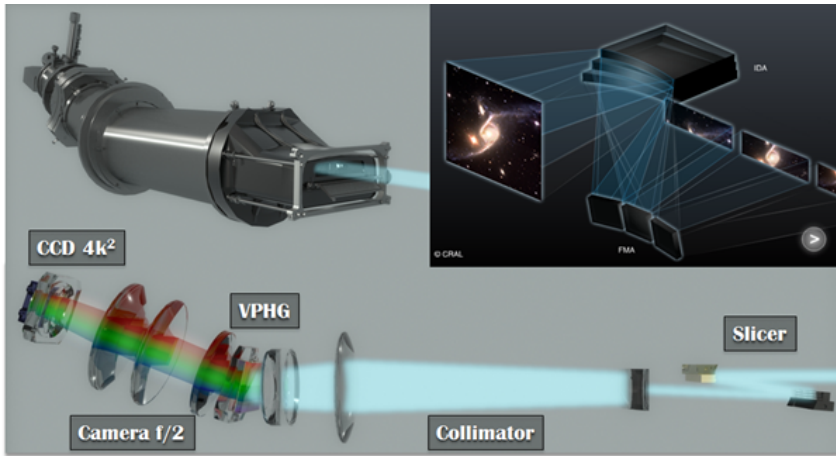


Fig. 2. Acquisition process of MUSE. The observed field is split in 24 IFU and each IFU dispatch the light on 48 slices. Image from <http://muse.univ-lyon1.fr> ©CRAL

2.2 Data Reduction Pipeline

Operations such as bias, darks and flat corrections are performed on each slice of data obtained from the CCD sensors. The slices are then merged together in a "pixel table" which contains spectral and spatial positions for each data pixel. Further operations are applied by modifying the coordinates, thus allowing to avoid any interpolation and resampling.

As MUSE is based on Earth-ground, the captured light has to go through the atmosphere. Analyzing faint remote objects spectra requires that the contribution of the atmosphere is taken into account and removed. The sky spectrum is estimated and removed using a method named ZAP developed for MUSE data detailed in (Soto et al., 2016).

Exploring the deep Universe requires to accumulate numbers of hours of observation in order to collect enough photons. However as cosmic rays may randomly blast detectors, single exposures are limited in time thus allowing to reject these cosmic rays as outliers. For example the MUSE observation of the Hubble-Deep-Field-South (HDFS) is composed of more than 50 half-hour long exposures for a total of about 25 hours of exposition time. Each exposure corresponds to a pixel table.

The last step consists in merging the single exposures using a sigma-clipping sum. It is at this stage that resampling is performed to align all exposures on a same grid before merging them. The result is a 300 by 300 pixels by 3600 wavelengths data cube associated with a variance cube of same dimensions. This latter is estimated by propagating the error estimated at the captor level at each stage of the processing.

These resulting data and variance cubes are the input of the methods developed

in the following.

3 Clustering approach

The approach chosen here to tackle the problem of the detection of the CGM is based on classical machine learning techniques. In a spectral narrow band around the Lyman emission, the studied galaxy and the possibly present halo are expected to have very similar spectral signature, whereas other pixels contain either noise spectra or different galaxies spectra that do not exhibit Lyman- α signature in this spectral band. By doing a 2 classes clustering in the spectral narrow band, the goal is to get a Lyman- α emitter pixel class and a noise pixel class. It can be outlined that the presence of two spatially very close galaxies (over neighbouring pixels, or even overlapping ones) is unlikely. However, in such a case, it is expected that these two objects are well separated in the physical space, meaning that they are at quite different distances from earth. Consequently, their respective spectra are well separated due to redshift.

3.1 Preprocessings

3.1.1 Variance reduction

Before any other processing of the data cube is performed, this latter is "reduced" by the variance cube, in order to lessen the influence of abnormal pixels.

For each pixel spectrum element (or voxel) $x_{i,\lambda}$, where i is the spatial position of the pixel and λ the spectral coordinate, one gets the reduced voxel $\tilde{x}_{i,\lambda}$ via normalizing by the variance voxel $v_{i,\lambda}$:

$$\tilde{x}_{i,\lambda} = \frac{x_{i,\lambda}}{\sqrt{v_{i,\lambda}}}$$

Each voxel is now a realization of a random variable of unitary variance.

3.1.2 Removing the continuum

Whereas halo are expected to emit narrow band Lyman- α spectra, galaxies can emit in all wavelengths and often in a brighter way than the halo emission line. Therefore a nearby galaxy can have a bright continuous spectral emission in the same band than the studied Lyman- α emission line. To avoid false detection due to this continuum spectrum, it is necessary to estimate and remove the continuum part in the spectrum of the analyzed pixels. By doing so, pixels containing a halo emission line spectra will be preserved and pixels of nearby galaxies will be reduced to noise.

A classical approach to estimate the (slowly varying) continuum part of a spectrum is the windowed median filter. A key parameter is then the window size which must be chosen at least twice the width of an emission line in order to smooth out this latter. The median filter has the advantage of being fast and

easy to implement but it can lead to important Mean Square Error (MSE) for the estimation of the continuum. It must be noted that for pixels with halo signal, overestimating the continuum part (due e.g. to the presence of an insufficiently smoothed out emission line) will result in a loss of the emission line signal after subtraction. As the SNR is already low, such a loss has to be avoided as much as possible.

In order to improve the continuum subtraction methods, a new approach based on a robust estimation of the continuum part has been described in (Bacher et al., 2016). By considering the emission peaks as outliers, it is proposed to compute a local robust linear regression to estimate the baseline (the continuum), without being biased by the peaks. The robust regression is based on the Least Trimmed Squares (LTS) algorithm that only takes into account the k smallest residuals, where k is the estimated number of non-outliers. The main contribution in (Bacher et al., 2016) is to propose a method for estimating the number k of non-outliers on the fly, in a given spectrum.

3.1.3 Matched Filter

To improve the Signal-to-Noise-Ratio(SNR), a filter matched to the point spread function (PSF) of the instrument is applied. The PSF is assumed to be separable in both spatial spread function (Field Spread Function or FSF) and spectral spread function (Line Spread Function or LSF). The key assumption here is that the main part of the noise is due to the instrument and it is therefore not convolved with the PSF. Although the searched object here is an extended halo and not a punctual source such as a galaxy (the image of a distant spatially unresolved galaxy matches precisely the FSF), this approach allows to significantly improve the SNR of halos. A drawback of applying the FSF matched filter is that it will slightly extend the signal spatially.

3.2 Unsupervised Clustering

As each Lyman- α galaxy emitter can be slightly different from the other ones, the first approach consists in applying an unsupervised clustering algorithm without any *a priori*.

A small spatial patch is defined around a galaxy and for each pixel, the narrow band spectrum is taken as the features vector.

3.2.1 Choice of a metric

The explored metrics were the euclidean distance, the Spectral Angular Distance (SAD) and the Spectral Information Divergence (SID)(Chang, 1999), defined as follow: For $\mathbf{x} = x_1 \dots x_n$, and $\mathbf{y} = y_1, \dots y_n$

- $d_{EUC} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$

- $d_{SAD} = \cos^{-1} \left(\frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}} \right)$
- $d_{SID} = \sum_{i=1}^n x_i \times \log\left(\frac{x_i}{y_i}\right) + y_i \times \log\left(\frac{y_i}{x_i}\right)$ avec $\forall i, x_i > 0$ and $y_i > 0$

Histograms of the different distances between the spectra of the pixels in the patch and a reference spectrum are shown in Figure 3. The reference spectrum is estimated at the center of the galaxy. After some tests, the SAD metric associated with the spectral clustering approach was retained. SAD is actually less sensible to the intensity of spectra than it is to the relative intensity distribution over the considered narrow band. This fits the requirement that the narrow band clustering should be essentially based on the form of the spectral peak. Furthermore SAD behaves like a normalized correlation between the spectra, which is especially useful as the dynamic is quite high between the galaxy and the halo. Note that SID divergence could have been a good candidate but it is limited to positive valued signals. Considered spectra are too noisy to be considered as such once data reduction and base line subtraction were applied.

3.2.2 Spectral clustering

Spectral Clustering (Ng et al., 2002) relies on an affinity matrix associated with a weighted graph. Here, a node \mathbf{x}_i is a spectrum of λ wavelengths (depending on the case it can be a narrow band spectrum or a large band one), associated with a pixel of an image centered on the studied galaxy.

Take \mathbf{M} a distance matrix associated with the graph defined as : for each node \mathbf{x}_i and \mathbf{x}_j of the graph, $M_{i,j} = d(x_i, x_j)$. A similarity matrix \mathbf{S} can for example be derived from the distance matrix by $S_{ij} = \exp^{-\frac{M_{ij}^2}{\sigma^2}}$ with σ being a key parameter assimilated with the characteristic size of a cluster. The normalised Laplacian matrix \mathbf{L} of the graph is then defined as

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}} \mathbf{S} \mathbf{D}^{-\frac{1}{2}}$$

with \mathbf{D} being the degree matrix : $D_{ij} = \sum_{j=1}^n S_{ij}$.

Spectral clustering consists in finding the k first eigenvectors of the Laplacian, k being the number of clusters, and projecting the data on the space spanned by these eigenvectors. Any unsupervised classification algorithm can then be used in the reduced k dimensional space. In the example below, a K-means algorithm was applied to identify the k -clusters.

3.2.3 Estimation of σ

As mentioned before, choosing the value of the parameter σ used to construct the similarity matrix is crucial. Many strategies may be tested for automatically

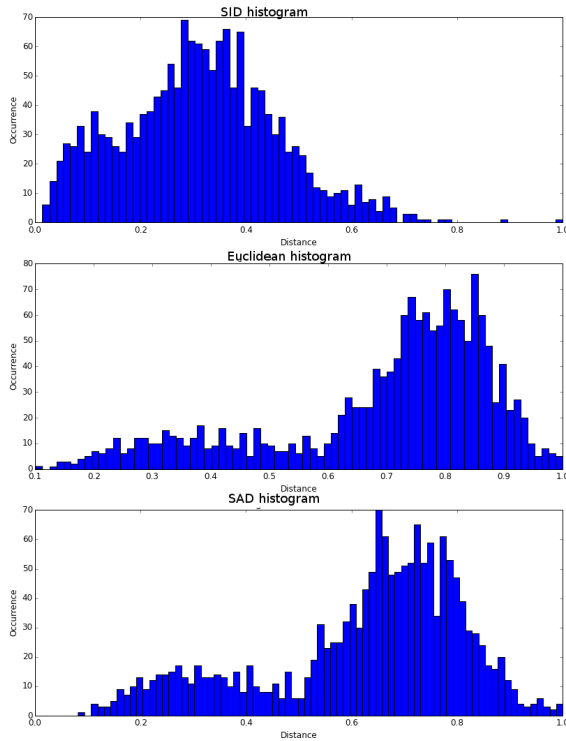


Fig. 3. Comparison of histograms for the different distances between galaxy neighborhood spectra and a reference spectrum estimated at the center of the galaxy

estimating σ . Minimizing the ratio between the intra-cluster variance and the extra-cluster variance seems a natural approach but it turned out to give too much importance to the pixels of the galaxy core, that have a highest SNR and the purest signature, and thus leads to reject all other pixels. Considering back the role and the physical interpretation of σ as the parameters that governs the characteristic size of a cluster, it has been tuned to match the mean of the intra-cluster distances. As it can be seen on the histogram of SAD distances shown in Figure 3, the histogram can be approximately decomposed into a mixture of two distributions. The first one is a single Gaussian distribution of distances (associated to the distances between the reference spectrum and a "random" background spectrum) and the second one is a distribution with a smaller mean associated to a set of correlated data (low values of SAD) that are likely to correspond to the pixels of interest.

The estimation of the Gaussian parameters of the first mixture component is performed by an Expectation Maximisation (EM) algorithm where the first step consist in estimating the probability of a pixel to belong to the Gaussian process, and the second step estimates the Gaussian parameters. Although this step already

performs a clustering, it must be noted that its performances are rather poor. This result is however kept for tuning the parameters of the more effective spectral clustering method. Although the histogram represents the distance to a reference spectra and not the intra-cluster distances, it makes sense to consider that the intra "noise" cluster distance is of the order of magnitude of the estimated standard deviation of the Gaussian detected mode. σ is set to match this latter value.

It can be noted that a semi-supervised approach based on (Wang and Davidson, 2010) was also explored, to try circumvent the problem of estimating σ . The idea is to add a number of constraints on known pair of pixels that must or cannot be part of a same cluster. Whereas this approach seems appealing as it allows to add additional already known informations; in practice results were no more satisfying than the unsupervised approach, as a really high proportion of pixels had to be "labelled" in order to find a good enough classification.

3.3 Postprocessing

To improve the robustness of the classification, a Potts model based spatial regularization step is proposed, implemented by a simulated annealing process. The Potts model (Pony et al., 2000) involves two energy contributions; one (U_{ext}) is related to the interaction between pixels and is governed by a tuning parameter β that acts as a temperature, and the second one (U_{int}) is the eigen energy associated to each individual pixel. We search to find the labels configuration c of the studied pixels that minimize the total energy U or equivalently that maximize the probability $P(C = c) = \frac{e^{-U(c)}}{Z}$ with Z the normalization term.

To set it precisely in our context, for each pixel the eigen energy term is given by euclidean distance between the pixel and the center of its class, and the regularization (or interaction) term is set to be the number of neighbors (e.g. in a 8-neighborhood) that does not possess the same label.

Take N the number of pixels, $\{c_i\}_{i=1..N}$, the configuration of the labels $\{l_i\}_{i=1..N}$ of the pixels $\{i\}_{i=1..N}$ resulting from the clustering and k the number of clusters (2 here). Then for $j \in \llbracket 0, k \rrbracket$,

$$P(l_i = j | c_i) = \frac{e^{-(U_{int}^{\setminus i} + \beta U_{ext}^{\setminus i} + U_{l_i=j}^i)}}{Z}$$

with $U_{\setminus i}$ the energy term summed on all pixels except i .

As the choice of β is an important issue, a marginalization of β is proposed, assuming a Gamma prior for β now seen as a random variable. In order to do this marginalization, Z , now function of β , is approximated by $Z(\beta) \simeq \beta^{-1}$. It follows that the local probability conditionally to β is approximated by

$$P(c = l_i | \beta) = \begin{cases} \beta e^{-\beta U_{ext}(x=l_i)} & \text{if } U_{ext}(c = l_i) \neq 0 \\ \beta e^{-\frac{\beta}{2}} & \text{elsewise} \end{cases}$$

Using a Gamma prior of parameters α_0, α_1 (fixed to 1 here), for β , and Bayes

formula, β can be marginalized and it comes that

$$P(c = l_i | C) \propto e^{-\frac{U_{int}(C, c=l_i)}{T}} \left(\sum_{c, U_{ext}(c=l_i) \neq 0} U_{ext}(c = l_i) + \frac{N_0}{2} + \frac{1}{\alpha_1} \right)^{-\frac{N+\alpha_0}{T}}$$

with T the annealing temperature, N the number of pixels and N_0 the number of homogeneous cliques, i.e. where $U_{ext}(c = l_i) = 0$.

The minimisation of the total energy of all pixels is then performed by a simulated annealing algorithm using this formulation of the *a posteriori* configuration probability.

Result of regularization using a marginalized β is shown in Figure 3.3.

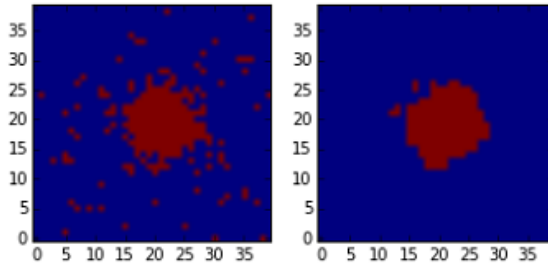


Fig. 4. Classification of one halo+galaxy without and with regularization with marginalized β

4 Results

4.1 MUSE data

The Hubble Deep Field South (HDFS), shown in Figure 5 is one of the deepest image of the Universe ever taken by the spatial telescope Hubble and as such was one of the first targeted by MUSE. This part of sky will be referred to as HDFS in the sequel. The details of the data reduction pipeline for MUSE observation of HDFS is described in (Bacon et al., 2015), along with the first galaxy spectral analysis results. To test the different halo detection approaches several objects of interest were chosen in the HDFS data cube, some with expected halo and some without expected halo. A 3D patch of 40 by 40 pixels by 30 wavelengths was selected, spatially centered on the object of interest and spectrally centered on the position of the Lyman emission peak.

4.2 Process pipelines

Pipelines are studied with different preprocessing choices. The estimation and removal of the continuum is performed by the LTS method developed in (Bacher



Fig. 5. HDFS image

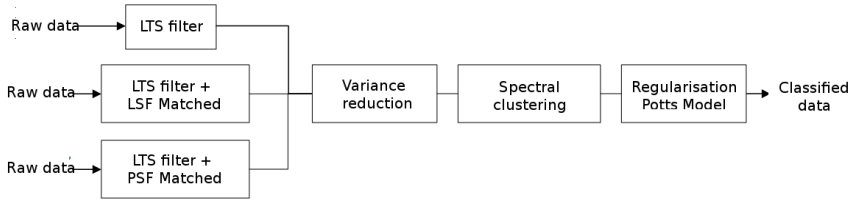


Fig. 6. Summary of process pipeline

et al., 2016) using a sliding window of width 100. Three cases are considered , without Matched Filter, with LSF based Matched Filter LSF or with a Matched Filter accounting for the whole PSF (LSF and FSF). The clustering steps are then identical for all three options, based on spectral clustering as described earlier. Finally a Potts model based spatial regularization with marginalization of β is applied.

4.3 Analysis

Results of the three different pipelines on the 9 chosen galaxies from HDFS are shown in Figure 4.3. The most striking result is that using the PSF-Matched Filter (spectral and spatial) clearly gives the most expanded detection. As outlined earlier, this result is of course biased by the spreading effect of the FSF. Nonetheless, it reveals peculiar asymmetrical structures surrounding several of the galaxies.

The importance of the continuum removal pre-processing step is stressed on the example of the 6th object: one can see that the galaxy of interest is really spatially close to other galaxies that appeared quite similar with respect to SAD

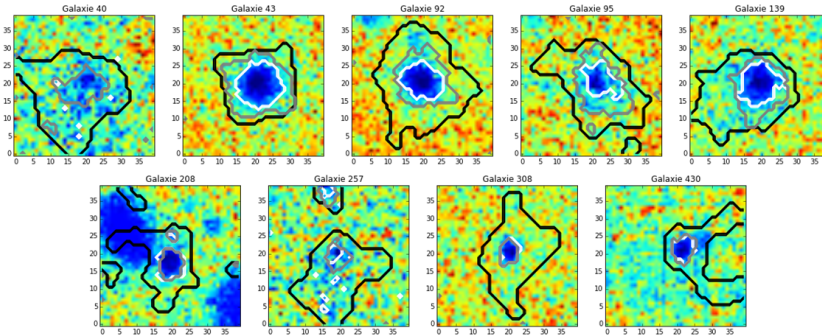


Fig. 7. Classification contour for each pipeline. Result without Matched Filter are in white, those obtained using LSF-Matched Filter are in gray and those with PSF-Matched Filter are in black. Represented underneath is the SAD distance to the center of the galaxy (before any preprocessing). Blue color means values close to 0 and therefore indicates spectra close (in the sense of SAD) to the spectrum of the galaxy center. Galaxies ID refer to the catalogue in (Bacon et al., 2015)

metric (they appear in blue). The reason is that they exhibit strong continuum spectral values in the Lyman band of interest. Preprocessing steps allow to reject these galaxies from the searched cluster. Consequently only the very peculiar extension from the galaxy core of interest seems to be detected.

5 Perspectives

These preliminary approaches highlight that some of the studied galaxies are most likely surrounded by a halo; very asymmetrical structures are revealed around some of these galaxies.

In order to improve the estimation of the halo spatial support, it seems necessary to take into account the spectral variability of the targeted halo; indeed the spectral signature shows higher variability with increasing distance to the galaxy center.

As such, raw clustering approaches may not be the best candidates. In particular we notice that it is rather difficult to derive a fully automatic approach for this problem, notably as the crucial σ parameter of the similarity matrix should be then tuned adaptively to account this variability. Up to our knowledge, such an approach has not yet been derived. Moreover classical constraints added within a semi-supervised approach did surprisingly not lead to improved results, and this remains to be investigated. Other paradigms have to be explored to be able to add easily some a priori knowledge.

Alternately, new methods are being developed, mostly based on hypothesis tests, that are dedicated to extend the halo detection as far as possible. An approach based on dictionary learning of the galaxy spectrum, and a maximum

correlation statistic test is currently developed by the authors.

One major difficulty in validating these methods is the absence of any ground-truth. This is why it is of great value to develop several different approaches to cross validate results. These different approaches will be benchmarked and compared on real MUSE data in collaborations with astrophysicists from CRAL.

References

- Bacher, R., Chatelain, F., and Michel, O.: 2016, in *IEEE International Conference on Acoustic, Speech and Signal Processing 2016*
- Bacon, R., Accardo, M., Adjali, L., Anwand, H., Bauer, S., Biswas, I., Blaizot, J., Boudon, D., Brau-Nogue, S., Brinchmann, J., et al.: 2010, in *SPIE Astronomical Telescopes+ Instrumentation*, pp 773508–773508, International Society for Optics and Photonics
- Bacon, R., Brinchmann, J., Richard, J., Contini, T., Drake, A., Franx, M., Tacchella, S., Vernet, J., Wisotzki, L., Blaizot, J., et al.: 2015, *A&A* **575**, A75
- Chang, C.-I.: 1999, in *Geoscience and Remote Sensing Symposium, 1999. IGARSS'99 Proceedings. IEEE 1999 International*, Vol. 1, pp 509–511, IEEE
- Courbot, J.-B., Mazet, V., Monfrini, E., and Collet, C.: 2016, in *IEEE International Conference on Acoustic, Speech and Signal Processing 2016*
- Meillier, C., Chatelain, F., Michel, O., Bacon, R., Piqueras, L., Bacher, R., and Ayasso, H.: 2016, *A&A* **588**, A140
- Ng, A. Y., Jordan, M. I., Weiss, Y., et al.: 2002, *Advances in neural information processing systems* **2**, 849
- Paris, S., Mary, D., and Ferrari, A.: 2013, *Signal Processing, IEEE Transactions on* **61(6)**, 1481
- Pony, O., Descombes, X., and Zerubia, J.: 2000
- Soto, K. T., Lilly, S. J., Bacon, R., Richard, J., and Conseil, S.: 2016, *Monthly Notices of the Royal Astronomical Society* **458(3)**, 3210
- Wang, X. and Davidson, I.: 2010, in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp 563–572, ACM
- Weilbacher, P. M., Streicher, O., Urrutia, T., Jarno, A., Pécontal-Rousset, A., Bacon, R., and Bohm, P.: 2012, in *SPIE Astronomical Telescopes+ Instrumentation*, pp 84510B–84510B, International Society for Optics and Photonics
- Wisotzki, L., Bacon, R., Blaizot, J., Brinchmann, J., Herenz, E., Schaye, J., Bouché, N., Cantalupo, S., Contini, T., Carollo, C., et al.: 2015, *arXiv preprint arXiv:1509.05143*