A&A 600, A54 (2017)
DOI: 10.1051/0004-6361/201630249
© ESO 2017

Astronomy & Astrophysics

The VIMOS Public Extragalactic Redshift Survey (VIPERS)

PCA-based automatic cleaning and reconstruction of survey spectra⋆

A. Marchetti[1,⋆⋆], B. Garilli[1], B. R. Granett[2,3], L. Guzzo[2,3], A. Iovino[2], M. Scodeggio[1], M. Bolzonella[4], S. de la Torre[5], U. Abbas[6], C. Adami[5], D. Bottini[1], A. Cappi[4,7], O. Cucciati[10,4], I. Davidzon[5,4], P. Franzetti[1], A. Fritz[1], J. Krywult[8], V. Le Brun[5], O. Le Fèvre[5], D. Maccagni[1], K. Małek[9], F. Marulli[10,11,4], M. Polletta[1,12,13], A. Pollo[9,14], L. A. M. Tasca[5], R. Tojeiro[15], D. Vergani[16], A. Zanichelli[17], S. Arnouts[5,18], J. Bel[19], E. Branchini[20,21,22], J. Coupon[23], G. De Lucia[24], O. Ilbert[5], T. Moutard[25,5], L. Moscardini[10,11,4], and G. Zamorani[4]

(Affiliations can be found after the references)

Received 14 December 2016 / Accepted 9 January 2017

ABSTRACT

Context. Identifying spurious reduction artefacts in galaxy spectra is a challenge for large surveys.
Aims. We present an algorithm for identifying and repairing spurious residual features in sky-subtracted galaxy spectra by using data from the VIMOS Public Extragalactic Redshift Survey (VIPERS) as a test case.
Methods. The algorithm uses principal component analysis (PCA) applied to the galaxy spectra in the observed frame to identify sky line residuals imprinted at characteristic wavelengths. We further model the galaxy spectra in the rest-frame using PCA to estimate the most probable continuum in the corrupted spectral regions, which are then repaired.
Results. We apply the method to ~90 000 spectra from the VIPERS survey and compare the results with a subset for which careful editing was performed by hand. We find that the automatic technique reproduces the time-consuming manual cleaning in a uniform and objective manner across a large data sample. The mask data products produced in this work are released together with the VIPERS second public data release (PDR-2).

Key words. galaxies: statistics – surveys – methods: statistical

1. Introduction

Large surveys of galaxy redshifts represent one of the primary means to explore the structure of the Universe and the evolution of galaxies. These include wide-angle surveys at relatively low redshift, notably the SDSS (York et al. 2000) and 2dFGRS (Colless et al. 2001), and deeper, narrower probes including VVDS (Le Fèvre et al. 2005; Garilli et al. 2008), DEEP2 (Newman et al. 2013), and zCOSMOS (Lilly et al. 2009; a more complete review of current and past surveys is presented by Guzzo et al. 2014). The recently completed VIMOS Public Extragalactic Redshift Survey (VIPERS) has built a sample that is simultaneously deep, densely sampled, and covers a volume similar to the 2dFGRS, but at $z = [0.5, 1.2]$ (Scodeggio et al. 2016; Guzzo et al. 2014; Garilli et al. 2014).

These surveys provide unique insight into both cosmology and galaxy formation. The spectra of extragalactic sources can enlighten our understanding of the underlying processes of galaxy evolution allowing measurements of physical properties such as star formation, metallicity, gas content and rotational velocity. The observed redshift contains not only information on the galaxy distance from the uniform Hubble expansion, but also the imprint of peculiar motions produced by the growth of structure. Statistically, the latter information can be extracted and used to test the nature of gravity (see e.g. the parallel papers by de la Torre et al. 2016; Hawken et al. 2017; Pezzotta et al. 2016). Knowledge of precise distances to galaxies also enables us to map the cosmic web and characterise galaxies with respect to the local density field; this is the starting point for studying the interplay between galaxy properties and their environment (as presented in the parallel paper by Cucciati et al. 2017).

Spectra obtained from ground–based optical surveys suffer from contamination from signal coming from sky emission: the Earth's atmosphere alongside instrumental and data reduction artefacts. This affects the identification and measurements of emission and absorption features, which are used to determine the redshift, and also corrupts estimates of line intensities, through which galaxy properties are characterised. These defects can be cured manually, when the number of spectra is small. In large modern surveys, however, automatic data reduction pipelines have become mandatory, as to efficiently manage the

⋆ based on observations collected at the European Southern Observatory, Cerro Paranal, Chile, using the Very Large Telescope under programs 182.A-0886 and partly 070.A-9007. Also based on observations obtained with MegaPrime/MegaCam, a joint project of CFHT and CEA/DAPNIA at the Canada-France-Hawaii Telescope (CFHT), that is operated by the National Research Council (NRC) of Canada, the Institut National des Sciences de l'Univers of the Centre National de la Recherche Scientifique (CNRS) of France, and the University of Hawaii. This work is based in part on data products produced at TERAPIX and the Canadian Astronomy Data Centre as part of the Canada-France-Hawaii Telescope Legacy Survey, which is a collaborative project of NRC and CNRS. The VIPERS web site is http://www.vipers.inaf.it/.

⋆⋆ Corresponding author: A. Marchetti,
e-mail: alida@lambrate.inaf.it

large quantity of data (e.g. Stoughton et al. 2002; Garilli et al. 2010, and references therein).

Spectroscopic data reduction pipelines perform the subtraction of sky lines and other known features; however, this operation is not free of errors, depending on various effects such as instrument optical distortions or the presence of fringing. As a result, after automatic sky subtraction, spurious residuals may be left and contaminate the spectrum, especially at $\lambda > 7500$ Å, where the sky emission becomes more and more dominant. For this reason, the processed spectra are usually inspected and often cleaned by human intervention, such as substituting the corrupted portion with a sensible interpolation. This cleaning pipeline facilitates and improves the quality of any spectral measurement performed on the calibrated spectrum such as the redshift, the line intensities, and spectral indices.

Unfortunately, repairing these defects is not always straightforward and in some cases may require the investment of a considerable amount of time. This is not feasible for the large numbers of spectra implied by the modern industry of redshift surveys: hundreds of thousands to millions of spectra are often collected within the same project. Additionally, such cleaning would be intrinsically subjective because different operators will apply different styles of cleaning across the same data, introducing some inhomogeneity. The only appropriate way to perform an efficient and objective cleaning of sky residuals or similar features from redshift survey spectra is then to implement a fully-controlled, automatic pipeline.

In this paper we describe the automatic pipeline we have developed within the VIPERS project. The algorithm identifies the position of residual artefacts that appear in the sky-subtracted spectra ("observed spectra" from now onwards) and creates a mask that matches their position; whenever possible, the algorithm reconstructs and repairs the corrupted spectral section. Both the identification and reparation of the affected spectral sections are based on the application of Principal component analysis (PCA; Karhunen 1947; Connolly et al. 1995; Yip et al. 2004).

A PCA-based sky subtraction was adopted by Wild & Hewett (2005) for the SDSS spectra, in which a set of sky emission templates was built by computing the principal components of the sky spectra, observed with a number of dedicated fibers. The closest sky contribution to each galaxy spectrum was then estimated and subtracted. Such an approach is appropriate for modeling and subtracting the sky spectra obtained with a fibre–fed spectrograph, but with VIPERS we face a different issue. VIMOS is a slitlet multi-object spectrograph, in which the sky is extracted from the fraction of slit adjacent to the object and then subtracted. This is done automatically by the data reduction pipeline (Scodeggio et al. 2005). All works well when the adjacent rows of sky spectrum are all aligned in wavelength with those containing the object spectrum. Unfortunately, optical distortions and fringing on the CCD surface break this symmetry by distorting sky emission lines along the slit, which leads to a sub-optimal subtraction that leaves residual features on the processed spectrum. Being related to the brightest sky features, these residuals appear at characteristic wavelengths.

The idea we have successfully developed in this paper has been to identify these residuals through a template spectrum, which is obtained by applying the PCA to the observed-frame galaxy spectra. In the observed frame, in fact, sky artefacts will sum up, while galaxy features will be, to some extent, suppressed because they appear at different redshifts. Given the stochastic nature of the residuals, however, this technique cannot be expected to reproduce the exact intensity and profile of each feature. Significant information will be contained in the high-order eigenvectors of the PCA in which less and less common details are encoded; these will get more and more mixed with real features from the galaxy spectra bacause many objects are at similar redshift, and thus share similar spectral structures in the observed frame.

We therefore have to limit ourselves to the first few eigenvectors (that we refer to as "eigenspectra"), in which the long redshift baseline of the survey guarantees that the main galaxy features are practically washed out. Under these conditions, the PCA reconstruction will not exactly reproduce the shape and intensity of the sky residuals, but can still be used to define a "mask" that marks the corrupted spectral ranges.

After the determination of the spectral sections to be masked we compute a realistic model of the spectrum to reconstruct the affected regions. This further step, unlike the masking procedure, requires the knowledge of galaxy redshifts. The contaminated regions we find are reconstructed by a second application of the PCA, this time performed in the galaxy rest frame (Marchetti et al. 2013).

Although designed for and calibrated on the VIMOS low-resolution spectra (refer to Scodeggio et al. 2016, for details), the pipeline is quite general and can be easily transported to other surveys.

The paper is structured as follows: in Sect. 2 we briefly introduce the data on which the method has been developed (VIPERS spectra), in Sect. 3 we make a general overview of the PCA method, in Sect. 4 we describe the residuals masking pipeline, in Sect. 5 its application to spectra, in Sect. 6 we describe the repairing of spectra within the masked regions and its advantages, and in Sect. 7 we summarize and draw the conclusions. In the Appendix we compare the automatic and manual masking of VIPERS spectra and discuss the results.

2. The VIPERS survey

The VIPERS survey spans an overall area of 23.5 deg^2 over the W1 and W4 fields of the Canada-France-Hawaii Telescope Legacy Survey Wide (CFHTLS-Wide). The VIMOS multi-object spectrograph (Le Fèvre et al. 2003) was used to cover these two regions through a mosaic of 288 pointings, 192 in W1 and 96 in W4. Galaxies to be targeted by VIPERS were selected from the CFHTLS catalogue to a limit of $i_{AB} < 22.5$, applying an additional $(r-i)$ vs. $(u-g)$ colour pre-selection that efficiently and robustly removes galaxies at $z < 0.5$. Coupled with a highly optimised observing strategy (Scodeggio et al. 2009), this doubles the mean galaxy sampling efficiency in the redshift range of interest, compared to a purely magnitude-limited sample, bringing it to 47%.

Spectra were collected at moderate resolution using the LR Red grism (7.14 Å/pixel, corresponding to an average $R \simeq 220$), providing a wavelength coverage of 5500–9500 Å. The typical redshift error for the sample of reliable redshifts (see below for definitions) is $\sigma_z = 0.00054(1+z)$; this corresponds to an error for the galaxy peculiar velocity at any redshift of 163 km s^{-1}.

The data were processed with the PANDORA EASYLIFE (Garilli et al. 2010) reduction pipeline. Redshifts and quality flags were measured with the PANDORA EZ (Easy Z) package (Garilli et al. 2010) and assigned to each spectrum. The redshift and flags (assigned by the PANDORA pipeline) were visually checked and validated, typically by two team members for each spectrum. The quality flag indicates the confidence level of the redshift measurement; for a description of the quality flag scheme see Scodeggio et al. (2016). In the PDR-2 VIPERS data

release, all the spectra are included with the exception of flag 0 spectra, which are objects without reliable redshift assignment (no identifiable features in the spectrum).

These are instead included in the analysis presented here, since they often provide the most information about artefacts from sky residuals, as they do not contain spectral features and carry information almost exclusively on the sky residuals contribution. Thus, this work used the whole set of 97 414 observed spectra, whose detail is given in Table 2 of Scodeggio et al. (2016).

Together with object spectra, the VIPERS survey database provides the sky spectra and the noise spectra associated to every observed spectrum (Garilli et al. 2010). The data from the final VIPERS Public Data Release (PDR-2) are available at online[1].

3. Principal component analysis on spectra

Principal component analysis (PCA) is a non-parametric way to reduce the complexity of high-dimensional datasets while preserving the majority of the information. PCA is possible when strong correlations exist in the data as is the case for galaxy spectra, which share many common features but yet are unique.

PCA consists of a linear transformation that changes the frame of reference from the observed, or natural, one to a frame of reference that highlights the structure and correlations in the data. The transformation aligns the principal axes with the directions of maximum variance in the data and is computed by diagonalising the data correlation (covariance) matrix. When applied to spectra containing flux measurements f_λ in M bins, the correlation matrix is given by

$$C_{\lambda_1,\lambda_2} = \frac{1}{N-1}\sum_{i=1}^{N} f^i_{\lambda_1} f^i_{\lambda_2}, \tag{1}$$

where i indexes the N spectra in the sample and λ_1 and λ_2 index the wavelength bins of the M^2 element correlation matrix.

The eigenvectors $e^i_{\lambda_j}$ of the sample, obtained diagonalizing the correlation matrix

$$C_{\lambda_1,\lambda_2} = \sum_{i=1}^{M} e^i_{\lambda_1} \Lambda_i e^i_{\lambda_2}, \tag{2}$$

represent the axes of the new coordinate system. The basis one obtains will be made up by orthogonal (i.e. uncorrelated) eigenvectors that are linear combinations of the original variables. The eigenvalues give the variance of the data in the orthogonal space and may be used to order the eigenvectors. By using only the most significant eigenvectors, meaning those corresponding to the largest eigenvalues, we can reconstruct most of the statistical information in the dataset.

Our data consist of N galaxy spectra each with M wavelength bins. Since the eigenspectra have the shape of spectra we refer to them as *eigenspectra*. When the spectra are kept in the observed frame, the signature of sky residuals is a coherent feature whereas the signal from astrophysical emission or absorption features is canceled out by the appearance of those features at different positions, due to the wide range of redshifts. Nonetheless, we still find a smooth signal representing the superposition of all galaxies continua, even if they are shifted with respect to each other according to their redshift. To eliminate this, we subtract the continua before computing the correlation matrix.

[1] http://vipers.inaf.it

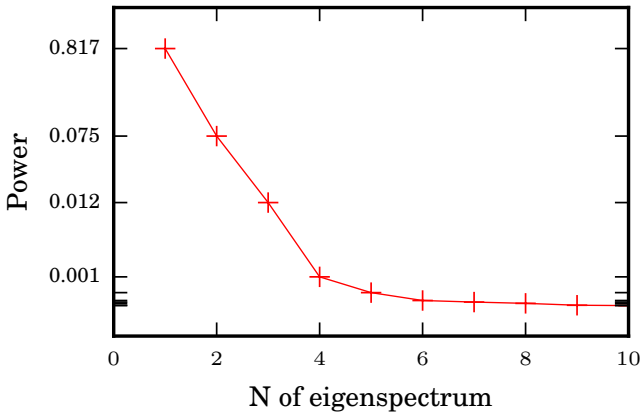

Fig. 1. Power associated with the first 10 eigenspectra. The labels on the vertical axis indicate the abscissae of the data points.

Each of the observed-frame spectral energy distributions, namely f_λ, can be expressed as a sum of the M eigenvectors with a set of M linear coefficients. We will truncate the sum to use only the first $K \ll M$ components:

$$\hat{f}_\lambda = \sum_{i=1}^{K} a_i e^i_\lambda, \tag{3}$$

where a_i are the linear coefficients. Projecting the observed-frame spectra onto the leading eigenspectra will give our best estimate of the locations and strengths of the sky residuals. We refer to the projection, $\hat{f}_\lambda$ in Eq. (3), as the "sky residuals spectrum". The choice of the number of components to take may be made based upon the relative power of each:

$$P(e_i) = \frac{\Lambda_i}{\Sigma^{\rm tot}_k \Lambda_k}, \tag{4}$$

where Λ_i stands for the ith eigenvalue, related to the ith eigenspectrum e_i.

4. The sky residuals eigenspectra

The first step of the method is to obtain the sky residuals eigenspectra. This is performed after subtracting the continuum and normalising the spectra by the scalar product. We estimate the continuum by convolving with a Gaussian kernel of width σ = 50 pixels, corresponding to 355 Å. After computing the eigenspectra, they are ordered with decreasing eigenvalue, such that the most common features within the spectra are contained in the first few eigenspectra.

To determine how many components to keep, we consider the size of the corresponding eigenvalues. Figure 1 shows the power associated with the first ten eigenspectra. The first eigenspectrum gives the average contribution of the sky residuals and it alone explains nearly 82% of the variance in the dataset. The second and third components encode 7.5% and 1.2% of the remaining information giving a total of 90% in the first three components. The information in the fourth component is significantly lower at 0.16% and the value of each higher order eigenspectrum decreases steadily. On the basis of this distribution of power as a function of the number of eigenspectra, we decided to use the first three to characterise the residual spectra. Using more does not improve the results and can lead to the over-fitting of astrophysical features. We thus construct a basis

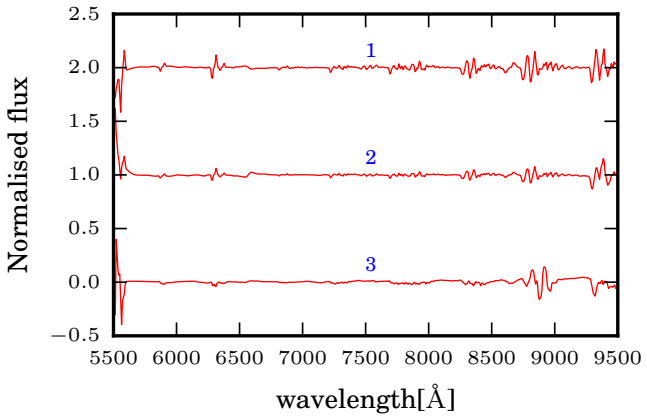

Fig. 2. Three principal components from the observed-frame VIPERS dataset. The first and second have been offset by two and one respectively, for visualisation convenience.

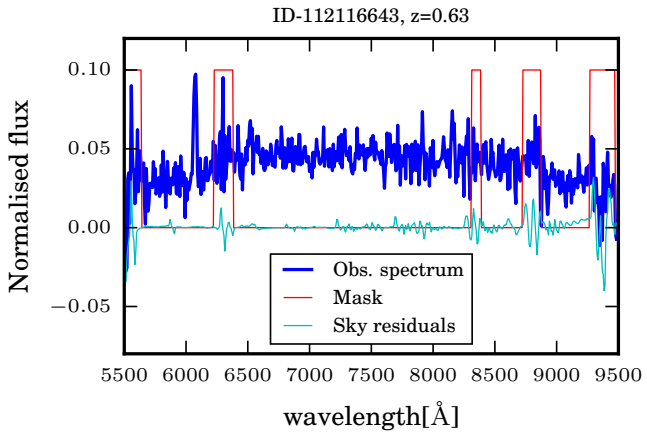

Fig. 3. Example of sky residuals spectrum (thin cyan) and its relevant observed spectrum (thick blue), from the VIPERS survey. The straight red lines indicate where the mask is applied on the basis of the sky residuals spectrum.

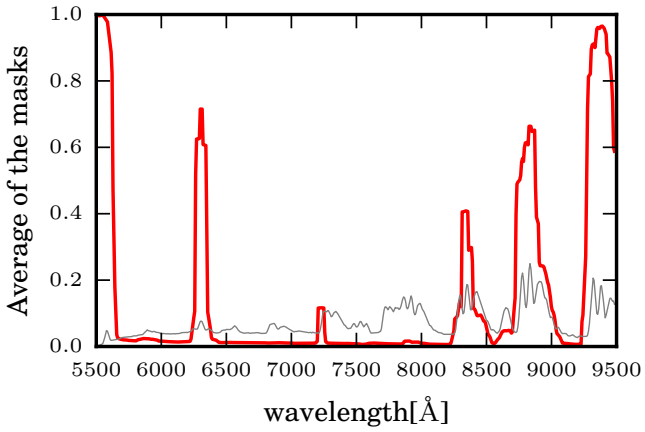

Fig. 4. Average of the masks for the VIPERS sample (thick red) with a VIPERS sky spectrum depicted in thin grey.

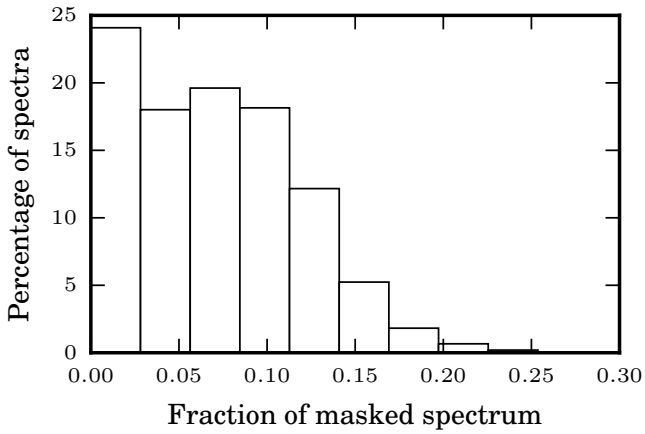

Fig. 5. Histogram for the distribution of the fraction of masked spectrum after the sky/zero-order residuals masking process. The regions at the edges of the spectra have been excluded.

from the three most significant eigenspectra, as shown in Fig. 2. The continuum-subtracted spectra are projected onto this basis to compute the residual spectra as shown in Fig. 3).

The main aim of this analysis is to define a mask for each spectrum, in correspondence with the more intense sky-residual features of the associated reconstructed sky residuals spectra. While the position of the sky residual features is recovered with reasonable accuracy, their intensity is often slightly over- or under-estimated as a consequence of using only few eigenspectra (see Marchetti et al. 2013). However, because the aim is to determine the position of the sky residuals, rather than to capture their precise strength, these discrepancies in intensity are not important.

5. Automatic masking of the spectra

We estimate the residuals of contamination for each galaxy spectrum, the "sky residuals spectrum", by projecting the spectra onto the basis of eigenspectra. The sky residuals spectrum is used to determine the threshold for masking. For each sky residuals spectrum, we compute the mean value and standard deviation σ. We mask all wavelengths where the sky residuals spectrum exceeds its mean by greater than $k\sigma$. The parameter k is set according to the characteristics of the dataset. For VIPERS we adopted 1.2σ at wavelengths shorter than 7500 Å, and 1.8σ at wavelengths longer than 7500 Å due to the higher contamination. These thresholds have been chosen empirically to select the known sky lines in a subset of representative spectra.

Figure 4 shows the frequency of masked wavelengths, obtained through the average of the masks, after applying the algorithm to the VIPERS dataset. Around 65% of spectra have been masked in correspondence to the λ = 6300 Å sky line and of the OH group at $\sim\lambda$ = 8700 Å; about 40% have been masked around λ = 8300 Å, and only the 10% at $\sim\lambda$ = 7300 Å. Nearly all spectra are masked at the upper and lower wavelength limits of the spectrum, where there is a significant contribution from fringing and calibration issues. In particular, at the lower limit there is the presence of the 5577 Å sky line residual combined with the fall-off in the detector sensitivity. The extent of the masking is summarised in Fig. 5. For half of the spectra, the mask covers less than 5% of the wavelength range.

Specifically for the VIPERS spectra, during this phase we also accounted for the extra residuals originating from the light coming from a bright object, lying on the slit adjacent to the one of the spectrum (zero-orders). An example of zero-order is shown in Fig. 6. These extra features were identified in the noise spectra delivered by the VIPERS reduction pipeline by undertaking a simple thresholding such the one applied for sky residuals,

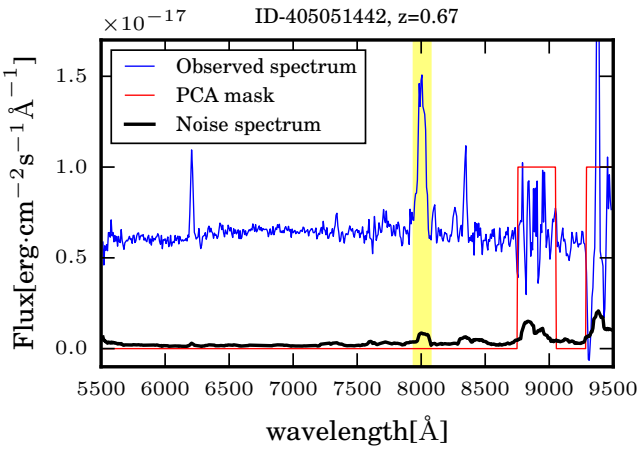

Fig. 6. Example of zero-order residual in a VIPERS spectrum (*top*, highlighted by the vertical yellow bar); at the *bottom* we show the corresponding noise spectrum. The clear excess corresponding to the zero-order position is used to supplement the sky residual mask (red line), as to account for this extra contribution in the final cleaning and repairing.

but with different empirical thresholds determined with a representative statistical subsample of spectra: we set the thresholds at 2σ, at wavelengths shorter than 7500 Å, and at 3σ at wavelengths longer than 7500 Å. Their position and size was thus added to the sky residual mask for subsequent treatment.

An additional check is then added to the pipeline, to ensure that the following step of the process (i.e. the repairing of masked regions, see Sect. 6) does not affect any relevant spectral features. Since the intensity of line features cannot be reproduced precisely by PCA, we introduce a "line safeguard" during the masking phase, by flagging the mask superposed to a feature with a different number with respect to the rest of the mask, to prevent using any subsequent PCA reconstruction at the locations of known features. For VIPERS we ensure that the reconstruction does not substitute the most prominent emission lines (e.g. [OII], Hβ, [OIIIa], [OIIIb], Hα for galaxies), or the D4000 break. For VIPERS data we chose the safety window as 28 Å to the sides of any emission line, and of 100 Å to preserve the D4000 Å break. These widths guarantee that the eventual repairing at the edges of this "particular" mask does not create an artificial feature mistakable for a physical one, where there was instead some artifact partially superposed to the feature itself. The safeguard was necessary for 30% of VIPERS spectra for which a mask fell on a known feature. The process described so far is summarised in the flow chart of Fig. 7.

6. Repairing the spectra

We next compute an estimate for the galaxy continuum in masked regions, to allow us to repair the contaminated data. This is helpful not only for visual inspection of the spectra but aids the measurement of spectral features as well. For example, line measurement tools require estimates of the continuum which may be unreliable due to spurious artefacts. Figure 8(top) shows the D4000 Å break for a VIPERS spectrum that is affected by an artefact that prevents the proper measurement of the intensity of the break.

Our approach to reconstruct and repair the spectrum is based on the PCA model in the rest-frame (Marchetti et al. 2013). The shift to rest-frame can only be made if the redshift is known; thus, we apply the pipeline only to galaxies with redshift quality

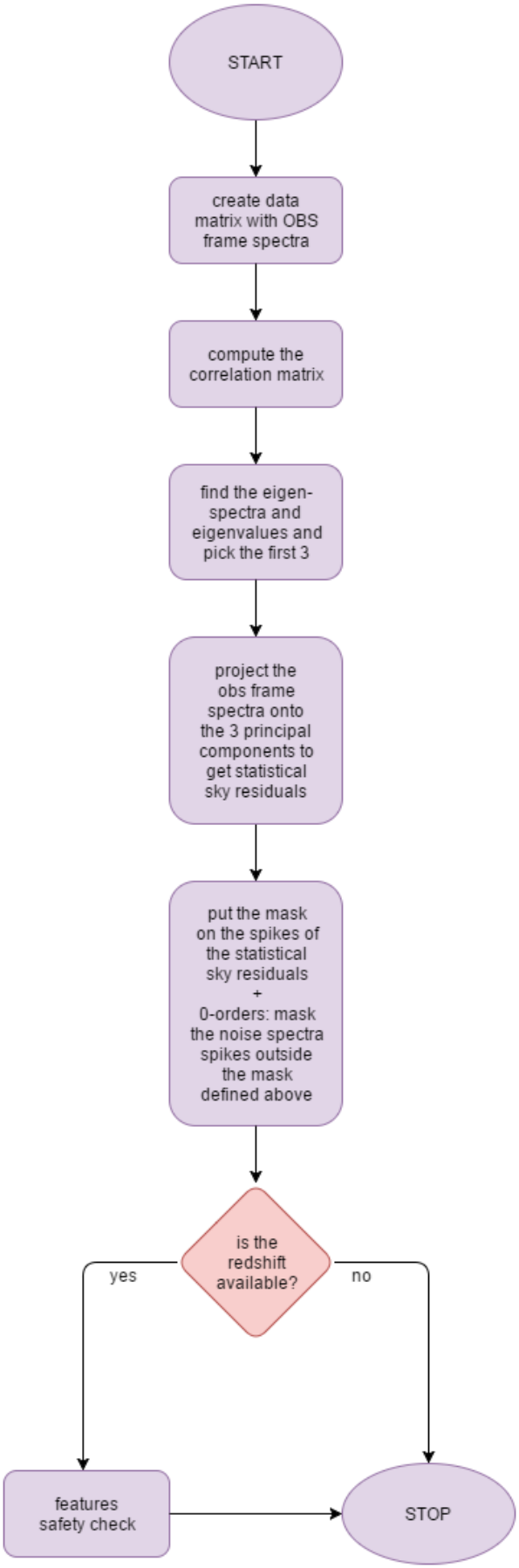

Fig. 7. Scheme of the automatic masking pipeline.

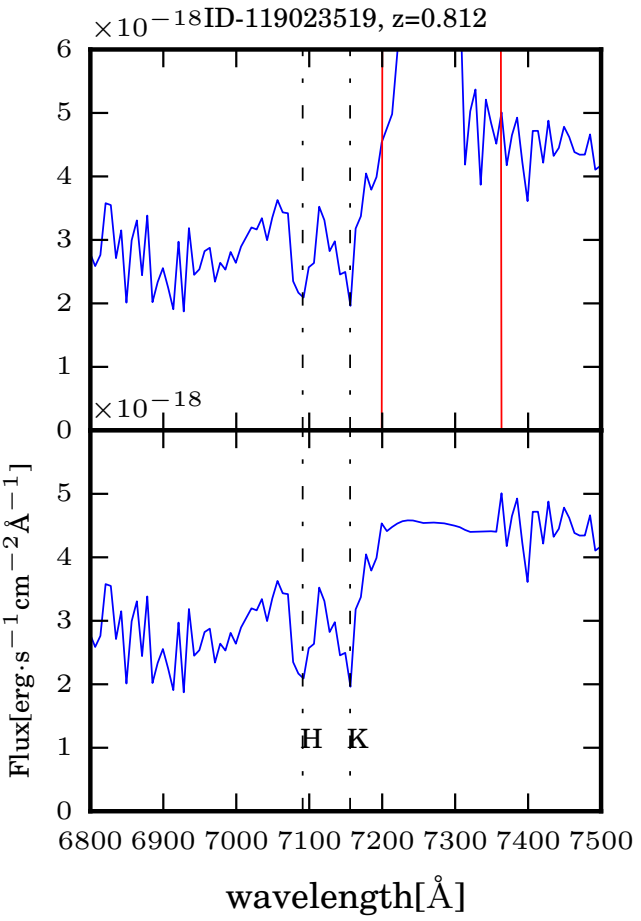

Fig. 8. *Top panel*: zoom of a VIPERS spectrum with strong residual at the right of the D4000 Å break. The region of the mask is delimited by the vertical red lines. *Bottom panel*: same VIPERS spectrum after repairing the portion affected by the residual.

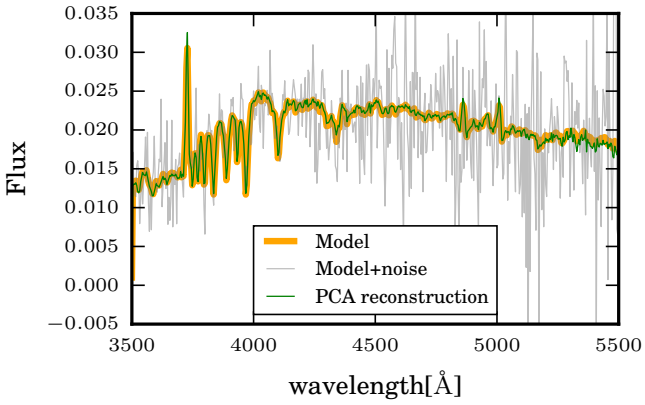

Fig. 9. Comparison between a model spectrum (thick orange) and the PCA reconstruction of the same spectrum (thin green) after degrading it with noise (in soft grey the degraded spectrum).

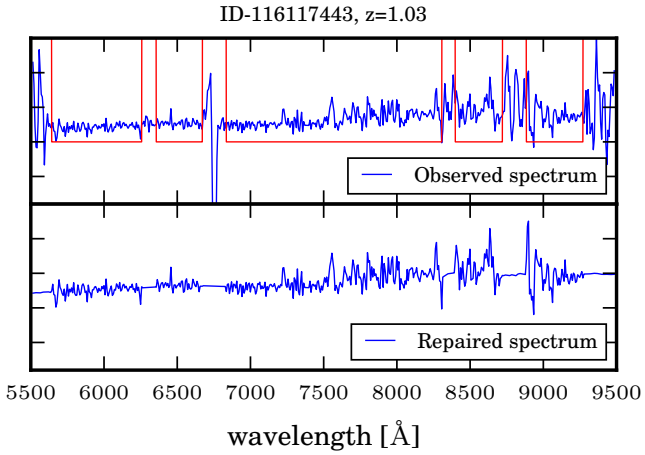

Fig. 10. Example of a VIPERS spectrum (blue) and the corresponding mask (straight red lines; *top panel*) and its PCA repairing in the masked portions (*bottom*).

flag >=1. Futhermore, the wavelength range in the rest-frame is limited by the redshift range of the sample. To have sufficient spectra we limit the sample to the redshift range $0.4 < z < 1.4$.

After shifting the spectra to the rest-frame, we compute the eigenspectra as described in Marchetti et al. (2013). We use the most significant three eigenvectors to reconstruct the spectra continuum.

The PCA with three components was found to accurately reconstruct the continuum of VIPERS spectra in Marchetti et al. (2013). Here we further demonstrate the accuracy of the reconstruction using mock spectra built from linear combinations of Bruzual-Charlot (Bruzual & Charlot 2003) and Kinney-Calzetti (Kinney et al. 1996) templates, which are also described in Marchetti et al. (2013). The mock spectra were redshifted and then degraded with Gaussian noise, to mimic the properties of the VIPERS spectra. Using a sample of 20 000 mock spectra we estimate the first three eigenspectra. We then project the mock spectra onto the basis of eigenspectra to compute the reconstruction. We plot an example spectrum in Fig. 9, showing excellent agreement between the model and the reconstruction at all wavelengths. While for the emission lines the discrepancies can be up to ~25% (Marchetti et al. 2013), the continua are always well reproduced. This is particularly the case for the test spectra, for which we found that the discrepancies between reconstructed and model continua are on average lower than ~1.6% and are never worse than ~5%, where such a discrepancy is obtained in rare cases (<0.1%).

An example of rest-frame repairing within the mask regions is shown in Fig. 10. After the repairing, the determination of the intensity of the galaxy spectral features is easier and more reliable, as shown in Fig. 8 bottom. This intensity can be quantified by running the EZ pipeline on the masked and repaired spectra: in measuring the intensity of spectral features, EZ associates an error to the spectral measurements. The natural consequence of the masking and repairing process is that the error associated with the spectral measurements is lower, since the noisy continuum adjacent to the spectral features is substituted with a highly reliable noiseless continuum (as demonstrated in Fig. 9). In particular for VIPERS we found that for about 64% of the spectra, the error associated to D4000 Å measurements is lower by ~11% on average for the masked and repaired sample. The same check, applied to the Hβ emission line, associates an error of ~30% lower to about the 54% of measurements.

Not all spectra may be repaired using this pipeline because sources outside the redshift range $0.4 < z < 1.4$ and/or sources without a measured redshift cannot be modeled using the PCA. For these sources we use a constant interpolation to repair the masked regions, as demonstrated in Fig. 11. Additionally, for active galaxies (AGN) the PCA reconstruction based upon galaxy

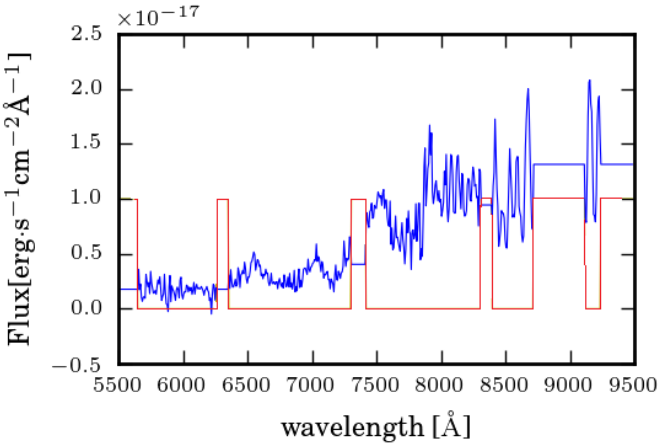

Fig. 11. Example of VIPERS stellar spectrum repaired with points at constant value at the level of the continuum, computed near to the regions of the mask (red).

eigenspectra is not applicable because Marchetti et al. (2013) found that the eigenspectra computed from the full survey could not represent rare AGN spectra well. Thus, for VIPERS spectra identified as AGN we have used constant interpolation to repair the masked regions. The steps followed to create the automatic repairing described here are schematically listed in the flow chart of Fig. 12.

7. Comparison and combination of automatic and manual cleaning of the VIPERS data

For VIPERS spectra, a significant amount of time has been spent by the VIPERS team to manually clean the spectra from sky residuals or other artefacts, producing many careful manual edits. To check the reliability and efficiency of the automatic pipeline, we compared ~500 automatically masked and repaired spectra with their corresponding manually edited spectra.

Figure 13 shows this comparison for two spectra of different quality. Overall, the automatically cleaned and repaired spectrum is very similar to the manually edited one. In the region of strong OH lines, the PCA cleaning looks more aggressive because the reconstructed portion of the spectrum is noise free while the human selectively edited the spurious features.

The bottom group of panels of Fig. 13 is like the previous, but shows a spectrum with lower signal to noise. In this case, the automatic cleaning is more precise with respect to the manual cleaning, especially around the 6300 Å sky line and the zero-order spectrum at 8700 Å.

8. Conclusions

We present a novel algorithm based upon principal component analysis to identify and repair spectral defects, such as those deriving from a non-perfect sky subtraction, in large sets of galaxy spectra. We have implemented this pipeline for the VIPERS dataset and tested its performance extensively against conventional manual spectral masking. The data products produced by this work are part of the VIPERS second data release (PDR-2; Scodeggio et al. 2016).

The PCA algorithm characterizes a dataset with a compact set of components without specification of a model. These components can represent the signal of interest but may also describe

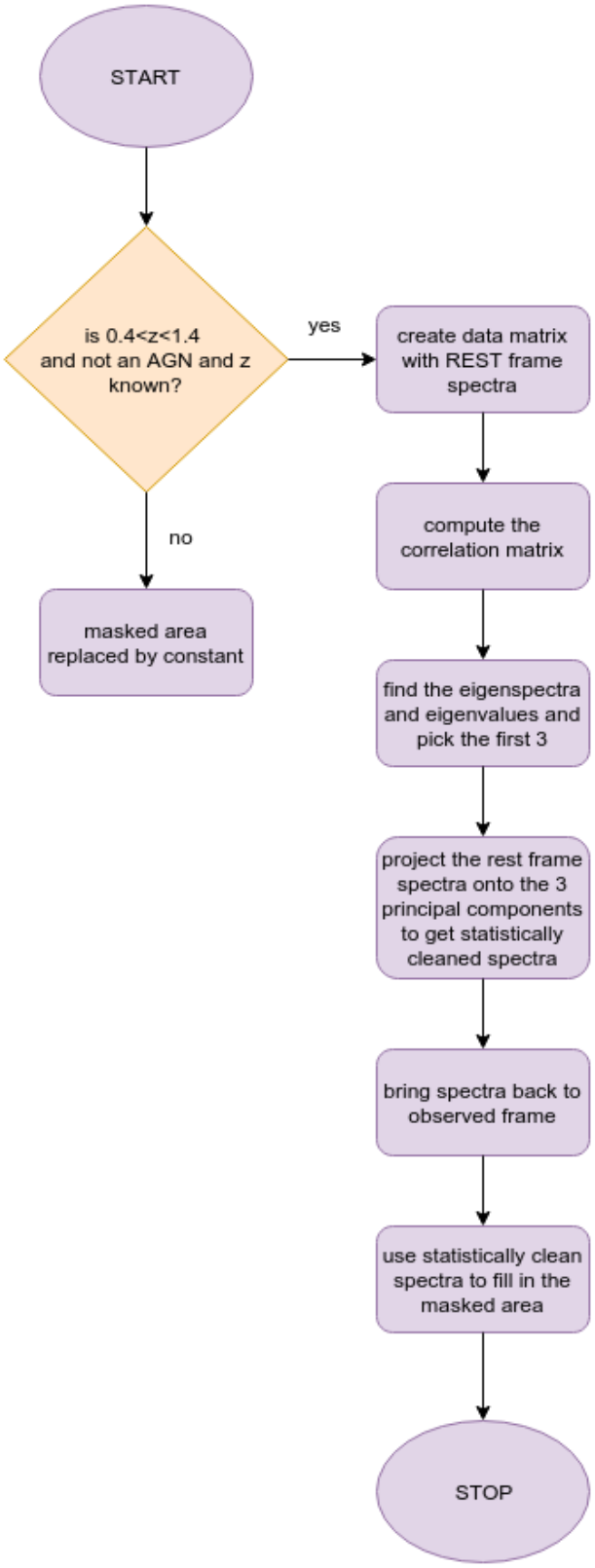

Fig. 12. Scheme of the automatic repairing pipeline.

unwanted systematic effects as we explored in this work. With the advent of spectroscopic surveys collecting millions of spectra, the use of automated pipelines is becoming unavoidable to guarantee the efficient and accurate treatment of the data. The pipeline described here has been tailored on VIPERS data, but it is general and can be easily applied to any large spectral survey,

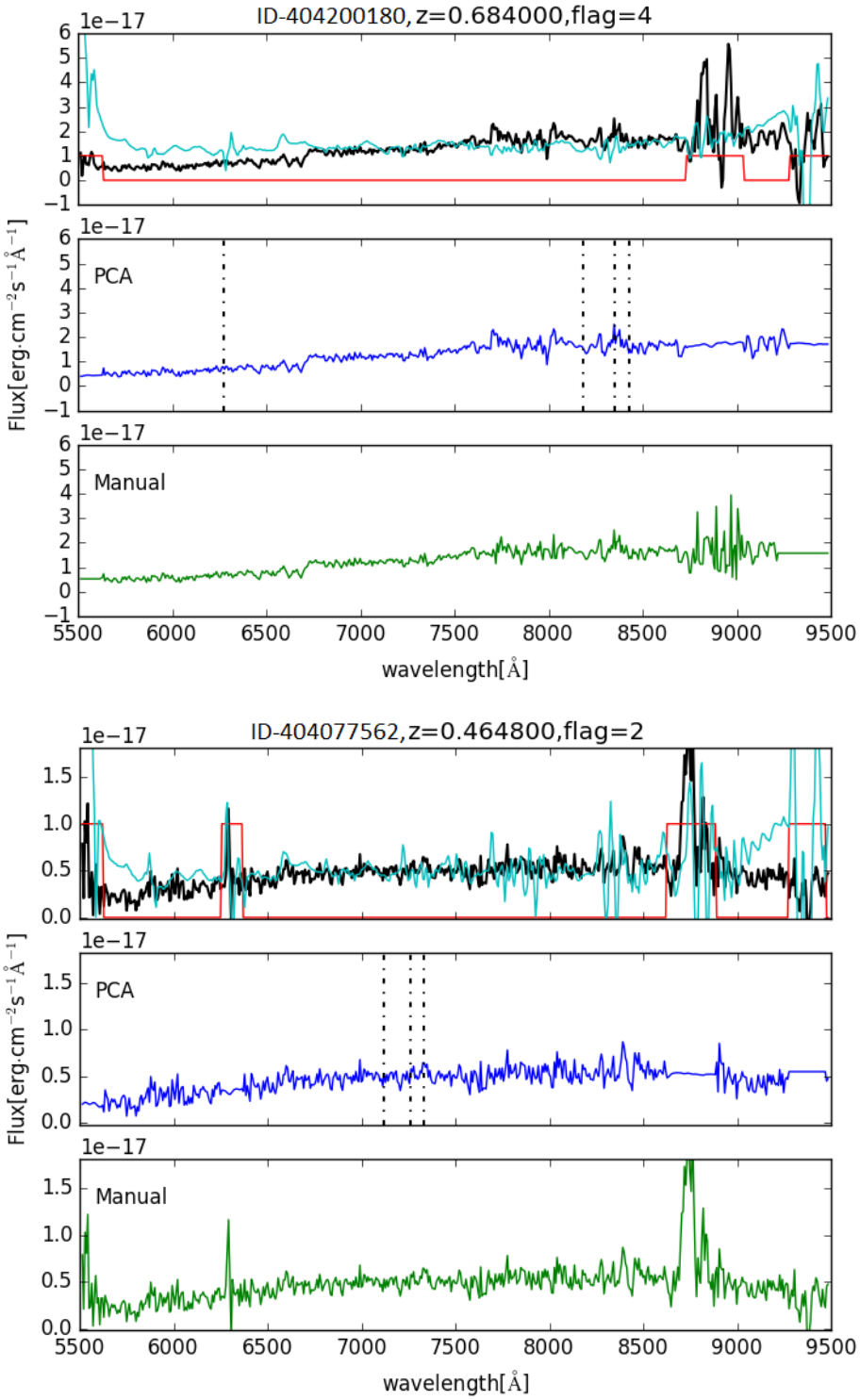

Fig. 13. Cleaning of a Flag 4 (*top*) and a Flag2 (*bottom*) VIPERS spectrum. For each panel: the *upper plot* is the observed (sky subtracted) spectrum (thick black), superposed to the mask (straight red lines) and the rescaled sky residuals spectrum (thin cyan); the *middle plot* shows the automatic cleaning, with the expected position of the [OII], Hβ and [OIII] lines marked in black by the dash-dotted lines; and the *bottom plot* is the manually edited spectrum.

after determination of the proper thresholds. The masking and repairing code is not available at present to the public but it can be modified and released in the future.

Acknowledgements. We acknowledge the crucial contribution of the ESO staff for the management of service observations. In particular, we are deeply grateful to M. Hilker for his constant help and support of this program. Italian participation to VIPERS has been funded by INAF through PRIN 2008, 2010, and 2014 programs. L.G., A.J.H., and B.R.G. acknowledge support from the European Research Council through grant No. 291521. OLF acknowledges support from the European Research Council through grant No. 268107. T.M. and S.A. acknowledge financial support from the ANR Spin(e) through the french grant ANR-13-BS05-0005. A.P., K.M., and J.K. have been supported by the National Science Centre (grants UMO-2012/07/B/ST9/04425 and UMO-2013/09/D/ST9/04030). W.J.P. is also grateful for support from the UK Science and Technology Facilities Council through the grant ST/I001204/1. E.B., F.M. and L.M. acknowledge the support from grants ASI-INAF I/023/12/0 and PRIN MIUR 2010-2011. L.M. also acknowledges financial support from PRIN INAF 2012. S.D.L.T. acknowledges the support of the OCEVU Labex (ANR-11-LABX-0060) and the A*MIDEX project (ANR-11-IDEX-0001-02) funded by the "Investissements d'Avenir" French government program managed by the ANR. and the Programme National Galaxies et Cosmologie (PNCG). Research conducted within the scope of the HECOLS International Associated Laboratory, supported in part by the Polish NCN grant DEC-2013/08/M/ST9/00664.

References

Bruzual, G., & Charlot, S. 2003, MNRAS, 344, 1000
Colless, M., Dalton, G., Maddox, S., et al. 2001, MNRAS, 328, 1039
Connolly, A. J., Szalay, A. S., Bershady, M. A., Kinney, A. L., & Calzetti, D. 1995, AJ, 110, 1071
Cucciati, O., Davidzon, I., Bolzonella, M., et al. 2017, A&A, in press, DOI: 10.1051/0004-6361/201630113
de la Torre, S. et al. 2016, A&A, submitted [arXiv:1612.05647]
Garilli, B., Le Fèvre, O., Guzzo, L., et al. 2008, A&A, 486, 683
Garilli, B., Fumana, M., Franzetti, P., et al. 2010, PASP, 122, 827
Garilli, B., Guzzo, L., Scodeggio, M., et al. 2014, A&A, 562, A23
Guzzo, L., Scodeggio, M., Garilli, B., et al. 2014, A&A, 566, A108
Hawken, A. J., Granett, B. R., Iovino, A., et al. 2017, A&A, accepted [arXiv:1611.07046]
Karhunen, H. 1947, Ann. Acad. Science Fenn, A.I, 37
Kinney, A. L., Calzetti, D., Bohlin, R. C., et al. 1996, ApJ, 467, 38
Le Fèvre, O., Saisse, M., Mancini, D., et al. 2003, in Proc. SPIE, 4841, eds. M. Iye, & A. F. M. Moorwood, 1670
Le Fèvre, O., Vettolani, G., Garilli, B., et al. 2005, A&A, 439, 845
Lilly, S. J., Le Brun, V., Maier, C., et al. 2009, ApJS, 184, 218
Marchetti, A., Granett, B. R., Guzzo, L., et al. 2013, MNRAS, 428, 1424
Newman, J. A., Cooper, M. C., Davis, M., et al. 2013, ApJS, 208, 5
Pezzotta, A., de la Torre, S., Bel, J., et al. 2016, A&A, submitted [arXiv:1612.05645]
Scodeggio, M., Franzetti, P., Garilli, B., et al. 2005, PASP, 117, 1284
Scodeggio, M., Franzetti, P., Garilli, B., Le Fèvre, O., & Guzzo, L. 2009, The Messenger, 135, 13
Scodeggio, M., Guzzo, L., Garilli, B., et al. 2016, A&A, submitted [arXiv:1611.07048]
Stoughton, C., Lupton, R. H., Bernardi, M., et al. 2002, AJ, 123, 485
Wild, V., & Hewett, P. C. 2005, MNRAS, 358, 1083
Yip, C. W., Connolly, A. J., Szalay, A. S., et al. 2004, AJ, 128, 585
York, D. G., Adelman, J., Anderson, J. J. E., et al. 2000, AJ, 120, 1579

[1] INAF–Istituto di Astrofisica Spaziale e Fisica Cosmica Milano, via Bassini 15, 20133 Milano, Italy
[2] INAF–Osservatorio Astronomico di Brera, via Brera 28, 20122 Milano, via E. Bianchi 46, 23807 Merate, Italy
[3] Università degli Studi di Milano, via G. Celoria 16, 20133 Milano, Italy
[4] INAF–Osservatorio Astronomico di Bologna, via Ranzani 1, 40127 Bologna, Italy
[5] Aix-Marseille Univ, CNRS, LAM, Laboratoire d'Astrophysique de Marseille, 13388 Marseille 13, France
[6] INAF–Osservatorio Astrofisico di Torino, 10025 Pino Torinese, Italy
[7] Laboratoire Lagrange, UMR 7293, Université de Nice Sophia Antipolis, CNRS, Observatoire de la Côte d'Azur, 06300 Nice, France
[8] Institute of Physics, Jan Kochanowski University, ul. Swietokrzyska 15, 25-406 Kielce, Poland
[9] National Centre for Nuclear Research, ul. Hoza 69, 00-681 Warszawa, Poland
[10] Dipartimento di Fisica e Astronomia – Alma Mater Studiorum Università di Bologna, viale Berti Pichat 6/2, 40127 Bologna, Italy
[11] INFN, Sezione di Bologna, viale Berti Pichat 6/2, 40127 Bologna, Italy
[12] Aix-Marseille Université, Jardin du Pharo, 58 bd Charles Livon, 13284 Marseille Cedex 7, France
[13] IRAP, 9 av. du colonel Roche, BP 44346, 31028 Toulouse Cedex 4, France
[14] Astronomical Observatory of the Jagiellonian University, Orla 171, 30-001 Cracow, Poland
[15] School of Physics and Astronomy, University of St Andrews, St Andrews KY16 9SS, UK
[16] INAF–Istituto di Astrofisica Spaziale e Fisica Cosmica Bologna, via Gobetti 101, 40129 Bologna, Italy
[17] INAF–Istituto di Radioastronomia, via Gobetti 101, 40129 Bologna, Italy
[18] Canada-France-Hawaii Telescope, 65–1238 Mamalahoa Highway, Kamuela, HI 96743, USA
[19] Aix-Marseille Univ, Univ Toulon, CNRS, CPT, 13453 Marseille, France
[20] Dipartimento di Matematica e Fisica, Università degli Studi Roma Tre, via della Vasca Navale 84, 00146 Roma, Italy
[21] INFN, Sezione di Roma Tre, via della Vasca Navale 84, 00146 Roma, Italy
[22] INAF–Osservatorio Astronomico di Roma, via Frascati 33, 00040 Monte Porzio Catone (RM), Italy
[23] Department of Astronomy, University of Geneva, Ch. d'Ecogia 16, 1290 Versoix, Switzerland
[24] INAF–Osservatorio Astronomico di Trieste, via G. B. Tiepolo 11, 34143 Trieste, Italy
[25] Department of Astronomy & Physics, Saint Mary's University, 923 Robie Street, Halifax, Nova Scotia, B3H 3C3, Canada