



Publication Year	2017
Acceptance in OA	2020-09-11T08:42:41Z
Title	QSFIT: automatic analysis of optical AGN spectra
Authors	CALDERONE, GIORGIO, NICASTRO, LUCIANO, GHISELLINI, Gabriele, Dotti, M., Sbarrato, Tullia, Shankar, F., Colpi, M.
Publisher's version (DOI)	10.1093/mnras/stx2239
Handle	http://hdl.handle.net/20.500.12386/27305
Journal	MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY
Volume	472

QSFIT: automatic analysis of optical AGN spectra

G. Calderone,^{1★} L. Nicastro,^{2★} G. Ghisellini,³ M. Dotti,⁴ T. Sbarrato,⁴ F. Shankar⁵
and M. Colpi⁴

¹INAF – Osservatorio Astronomico di Trieste, Via Tiepolo 11, I-34131 Trieste, Italy

²INAF – Istituto di Astrofisica Spaziale e Fisica Cosmica, Via Piero Gobetti 101, I-40129 Bologna, Italy

³INAF – Osservatorio Astronomico di Brera, Via E. Bianchi 46, I-23807 Merate, Italy

⁴Dipartimento di Fisica ‘G. Occhialini’, Università di Milano – Bicocca, Piazza della Scienza 3, I-20126 Milano, Italy

⁵Department of Physics and Astronomy, University of Southampton, Highfield SO17 1BJ, UK

Accepted 2017 August 30. Received 2017 August 29; in original form 2016 December 5

ABSTRACT

We present QSFIT (*Quasar Spectral Fitting* package), a new software package to automatically perform the analysis of active galactic nuclei (AGNs) optical spectra. The software provides luminosity estimates for the AGN continuum, the Balmer continuum, both optical and ultraviolet iron blended complex, host galaxy and emission lines, as well as width, velocity offset and equivalent width of 20 emission lines. Improving on a number of previous studies on AGN spectral analysis, QSFIT fits all the components simultaneously, using an AGN continuum model which extends over the entire available spectrum, and is thus a probe of the actual AGN continuum whose estimates are scarcely influenced by localized features (e.g. emission lines) in the spectrum. We used QSFIT to analyse 71 251 optical spectra of Type 1 AGN at $z < 2$ (obtained by the Sloan Digital Sky Survey, SDSS) and to produce a publicly available catalogue of AGN spectral properties. Such catalogue allowed us (for the first time) to estimate the AGN continuum slope and the Balmer continuum luminosity on a very large sample, and to show that there is no evident correlation between these quantities the redshift. All data in the catalogue, the plots with best-fitting model and residuals, and the IDL code we used to perform the analysis, are available on a dedicated website. The whole fitting process is customizable for specific needs, and can be extended to analyse spectra from other data sources. The ultimate purpose of QSFIT is to allow astronomers to run standardized recipes to analyse the AGN data, in a simple, replicable and shareable way.

Key words: methods: data analysis – catalogues – galaxies: active – quasars: emission lines.

1 INTRODUCTION

Statistical studies on large samples of active galactic nuclei (AGNs) have proven to be valuable tools to understand the AGN phenomenon (e.g. Khachikian & Weedman 1974; Kellermann et al. 1989; Sanders et al. 1989; Urry & Padovani 1995).

The ongoing and future large-scale extragalactic surveys require more appropriate, rapid and flexible tools to properly reduce and statistically analyse the vast amounts of already available and incoming data. For AGN, some examples of key quantities to extract from the data are the composite broad-band Spectral Energy Distributions (SEDs) (Elvis et al. 1994; Richards et al. 2006), composite optical/ultraviolet (UV) spectra (Francis et al. 1991; Vanden Berk et al. 2001; Telfer et al. 2002), or the compilation of catalogues of spectral properties (e.g. Shen et al. 2011). These works allow us

to highlight specific features of large samples, which are typically hidden when considering individual sources.

The data analysis procedures employed so far, even if meticulously described, have been very often based on ‘private’ algorithms, which do not allow other research groups to replicate the analysis, check for inconsistencies, customize the procedure for specific needs, or apply the analysis recipes to new data (e.g. Vestergaard et al. 2008; Shen et al. 2011). Nevertheless, the employed software is (at least) good enough to produce results worth a publication, therefore it is likely good enough to be publicly released.¹ Moreover, the availability of a well-tested software to perform data analysis may help to standardize the analysis recipes, which would otherwise be heterogeneously developed, and the results possibly hard to compare directly. In the field of X-ray data analysis, a very important example of such a standardization is the

★ E-mail: calderone@oats.inaf.it (GC); nicastro@iasfbo.inaf.it (LN)

¹ <http://www.nature.com/news/2010/101013/full/467753a.html>

XSPEC package,² which allowed researchers to focus on the interpretation of results, instead of dealing with the many technical details of the analysis procedures.

In the field of optical/UV AGN spectral analysis, there is not yet a commonly accepted and widely used software tool to estimate the relevant spectral quantities, namely the AGN continuum luminosities, slopes, emission-line widths, luminosities, equivalent widths, etc., although several works have tackled the problem of estimating these quantities. For instance the software by Calistro Rivera et al. (2016) requires data from infrared (IR) to X-rays and does not provide emission-line estimates. The works by Boroson & Green (1992) and Shen et al. (2011) provide the same estimates provided by QSFIT, but they did not release the software code. Finally, general purpose packages such as SPECFIT (Kriss 1994) provide just the basic fitting functionalities, but they do lack a complete analysis recipe to go from the spectrum in the FITS file to the final estimate of a spectral quantities.

In this work we present, and make publicly available to the community, our new and extensively tested software named QSFIT (*Quasar Spectral Fitting* package). In its first release (version 1.2), the software provides estimates of the spectral quantities by simultaneously fitting all components (AGN continuum, Balmer continuum, emission lines, host galaxy and iron blended lines) of the observed optical spectrum. The software is written in IDL, and it is released under the GNU General Public License (GPL) license (Section 2). It currently only operates on spectra observed by the Sloan Digital Sky Survey, Data Release 10 (SDSS-DR10³) and works non-interactively, in order to provide the simplest possible interface to the user. However, QSFIT can be customized for specific needs, and can be extended to analyse spectra from other data sources or other detectors, e.g. to perform a spectral analysis using both optical and UV data. The ultimate purpose of QSFIT is to allow astronomers to run standardized recipes to analyse the AGN data in a simple, replicable and shareable way.

In this work, we also present a catalogue of spectral quantities obtained by running QSFIT on a sample of 71 251 Type 1 AGN at $z < 2$ observed by SDSS-DR10. All data in the catalogue, as well as the plots with best-fitting model and residuals, the IDL code required to replicate all our data analysis, are available online.⁴ Our catalogue is similar to the one from Shen et al. (2011, hereafter S11), although the data analysis follows a different approach (Section 4). Moreover, we perform a comparison of our results with those from S11 in order to assess the reproducibility of the results in both catalogues, and discuss the differences. Finally, we check whether the AGN continuum slopes with QSFIT show any evolution with the redshift.

This paper is organized as follows: in Section 2, we present the QSFIT software in detail; in Section 3, we discuss the Type 1 AGN sample and the data analysis employed to build the QSFIT catalogue; in Section 4, we compare our results with those of S11; in Section 5, we discuss our results and in Section 6, we draw our conclusions. Finally, in Appendix B, we describe the procedures for QSFIT installation, usage and customization.

In this work, we adopt a standard Λ cold dark matter cosmology with $H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$, $\Omega_M = 0.3$ and $\Omega_\Lambda = 0.7$ (exactly the same as S11).

² <https://heasarc.gsfc.nasa.gov/xanadu/xspec/>

³ <https://www.sdss3.org/dr10/>

⁴ <http://qsfit.inaf.it/>

2 QSFIT: QUASAR SPECTRAL FITTING PROCEDURE

QSFIT (version 1.2) is an IDL⁵ package able to perform AGN spectral analysis at optical/UV wavelengths. It is ‘free software’,⁶ released under the GPL license, and available for download on Github.⁷

QSFIT operates on the optical spectrum of a single source and provides estimates of: AGN continuum luminosities and slopes at several rest-frame wavelengths; Balmer continuum luminosity; luminosities, widths and velocity offsets of 20 emission lines ($H\alpha$, $H\beta$, $Mg II$, $[O III]$, $C IV$, etc.); luminosities of iron blended lines at optical and UV wavelengths; host galaxy luminosities at 5500 Å (for sources with $z \leq 0.8$). Beyond the spectral quantities, the software also provides several ‘quality flags’ to assess the reliability of the results.

The purpose is similar to that of previous works on AGN spectral analysis, but we followed a different approach:⁸ instead of focusing on a single emission line and estimate the continuum in the surrounding region, we fit all the components simultaneously using an AGN continuum component which extends over the entire available spectrum, and is thus a probe of the actual AGN continuum whose estimates are scarcely influenced by localized features in the spectrum.⁹

The model used to fit the data is a collection of several ‘components’, each representing a specific contribution to the observed optical/UV spectrum. The list of components currently employed is:

- (i) AGN continuum (Section 2.2): we use a single power law to describe the AGN continuum over the entire (rest-frame) wavelength coverage (Cristiani & Vio 1990);
- (ii) Balmer continuum (Section 2.3): we follow the recipe by Grandi (1982) and Dietrich et al. (2002) to model the Balmer continuum. We also account for the high-order Balmer lines ($7 \leq n \leq 50$, i.e. $H\epsilon$ and higher) which are not fitted individually, using the line ratios given in Storey & Hummer (1995);
- (iii) Host galaxy (Section 2.4): we use the template of an elliptical galaxy component to account for the host galaxy contribution (this component is relevant only for sources with $z \leq 0.8$);
- (iv) Iron blended emission lines (Section 2.5): we used the iron template from Véron-Cetty, Joly & Véron (2004) and Vestergaard & Wilkes (2001) at optical and UV wavelengths respectively;
- (v) Emission lines (Section 2.6): we used a Gaussian profile to represent all the emission lines (both ‘narrow’ and ‘broad’) in

⁵ The QSFIT project started a few years ago, when the IDL language seemed the most viable option because of its widespread usage in the astronomy community. Today, PYTHON appears more appropriate, since it would allow QSFIT to be a complete open source solution. Hence, we started the PYTHON porting, which will however maintain most of QSFIT current architecture in order to facilitate the transition to PYTHON even to those who will start using QSFIT in its IDL version.

⁶ <https://www.gnu.org/philosophy/free-sw.html>

⁷ <https://github.com/gcalderone/qsfit>

⁸ See also Vestergaard et al. (2008) and Shen & Liu (2012).

⁹ In the few cases, where this approach is not appropriate (because of peculiar continuum shapes, strong absorptions or atypical iron emission patterns), the user can customize the source code to find a sensible fit. In our catalogue (Section 3), we excluded the sources with $z > 2$ (since their continuum cannot be fitted with a single power law) and the BAL sources (to avoid strong absorptions). The goodness of the fit for the remaining cases can be inspected by checking the χ^2 of the fit and the plots.

Table 1. The list of all the considered emission lines. The third column (*type*) tells whether the line is modelled with a broad (B) or a narrow (N) component. The fourth column (*Field name*) shows the prefix of the FITS column name for each considered emission line. The H α line is modelled with three components to account for the very broad feature below that line.

Line	Wavelength (Å)	Type	Field name
Si iv	1399.8	B	br_siiiv_1400
C iv	1549.48	B	br_civ_1549
C iii]	1908.734	B	br_ciii_1909
Mg ii	2799.117	B	br_mgii_2798
[Ne vi]	3426.85	N	na_nevi_3426
[O ii]	3729.875	N	na_oii_3727
[Ne iii]	3869.81	N	na_neiii_3869
H δ	4102.89	B	br_hd
H γ	4341.68	B	br_hg
H β	4862.68	B	br_hb
		N	na_hb
[O iii]	4960.295	N	na_oiii_4959
[O iii]	5008.240	N	na_oiii_5007
He i	5877.30	B	br_hei_5876
[N ii]	6549.86	N	na_nii_6549
H α	6564.61	B	br_ha
		N	na_ha
		(see the text)	line_ha_base
[N ii]	6585.27	N	na_nii_6583
[Si ii]	6718.29	N	na_sii_6716
[Si ii]	6732.67	N	na_sii_6731

the spectrum. Beyond the ‘known’ emission lines expected to be relevant in the considered wavelength range (see Table 1), we also considered a list of ‘unknown’ emission lines, i.e. priori not associated with any known line (Section 2.7). By considering these additional components, we may account for the lack of an iron template in the wavelength range 3100–3500 Å, or for asymmetric profiles in known emission lines.

The actual model is the sum of all the above-mentioned components. No absorption line is considered, since its parameters would be highly degenerate with either the host galaxy or the emission lines.

QSFIT relies on the MPFIT¹⁰ (Markwardt 2009) procedure as minimization routine: during the fitting process the component’s parameters are varied until the differences between the data and the model are minimized, following a Levenberg–Marquardt least-squares minimization algorithm.

Currently, QSFIT operates only on SDSS-DR10 spectra of sources whose redshift is $z < 2$ (to avoid the absorptions troughs in the neighbourhood of the hydrogen Lyman- α line). The whole analysis process is tuned to operate on the SDSS spectra and to generate the QSFIT catalogue (Section 3), but it can be modified to analyse spectra from other sources or to customize the fitting recipe (Appendix B). The details of the fit procedure are discussed in Section 2.8.

2.1 Units

The units and symbols used throughout the paper are as follows:

- (i) λ [Å]: wavelength;
- (ii) ν [Hz]: frequency;
- (iii) L_λ [10^{42} erg s $^{-1}$ Å $^{-1}$]: luminosity density;

(iv) L [10^{42} erg s $^{-1}$]: emission-line integrated luminosity, continuum νL_ν or λL_λ luminosities, etc.;

(v) α_λ AGN spectral slope, defined as $L_\lambda \propto \lambda^{\alpha_\lambda}$;

(vi) $\lambda_{\min}$ and $\lambda_{\max}$ [Å]: minimum and maximum rest-frame wavelength for a specific spectrum;

(vii) FWHM and V_{off} [km s $^{-1}$]: full width at half-maximum and velocity offset of emission lines. Both values are calculated as the width/displacement of the emission line, normalized by the reference wavelength, and multiplied by the speed of light. Positive values of V_{off} means the line is blueshifted.

All quantities are given in the source rest frame.

2.2 AGN continuum

The AGN broad-band continuum is modelled as a power law in the form:

$$L_\lambda = A \left(\frac{\lambda}{\lambda_b} \right)^{\alpha_\lambda} \quad (1)$$

where λ_b is a reference break wavelength, A is the luminosity density at $\lambda = \lambda_b$ and α_λ is the spectral slope at $\lambda \ll \lambda_b$. In the current QSFIT implementation, the A parameter is constrained to be positive; λ_b is fixed to be in the middle of the available spectral wavelength range and α_λ is constrained in the range $[-3, 1]$.

The AGN continuum component extends through the whole observed wavelength, hence the luminosity and slopes estimates are scarcely influenced by localized features in the spectrum, and are assumed to be reliable probes of the actual shape of the ‘real’ AGN continuum.

If the spectral coverage is sufficiently large the single power-law model may not be suitable to fit the data. In these cases, the continuum component can be customized to behave as a smoothly broken power-law model (see Appendix B5). For the QSFIT catalogue, we chose to use a simple power law, since the SDSS spectral coverage provides only weak constraints to a second power-law component. Note also that when the SDSS data are used and the host galaxy is significantly more luminous than the AGN continuum, the slope parameter becomes highly degenerate with the galaxy normalization, and the two parameters cannot be reliably constrained. This occurs typically for sources with $z \leq 0.6$ (Fig. 19), hence for these sources we fix the continuum slope to $\alpha_\lambda = -1.7$, i.e. the average value for the sources with $z \sim 0.7$ (see Fig. 21 and discussion in Section 5.3).

2.3 Balmer continuum and pseudo-continuum

The Balmer continuum (Wills, Netzer & Wills 1985) is modelled according to Grandi (1982) and Dietrich et al. (2002):

$$L_\lambda = A \times B_\lambda(T_e) \left\{ 1 - \exp \left[-\tau_{\text{BE}} \left(\frac{\lambda}{\lambda_{\text{BE}}} \right)^3 \right] \right\} \quad (2)$$

where A is the luminosity density at 3000 Å, in units of the continuum (Section 2.2) luminosity density at the same wavelength, $B_\lambda(T_e)$ is the blackbody function at the electron temperature T_e , τ_{BE} is the optical depth at the Balmer edge and λ_{BE} is the edge wavelength (3645 Å). To avoid degeneracies among parameters, we fixed the electron temperature at $T_e = 15\,000$ K and $\tau_{\text{BE}} = 1$ (Dietrich et al. 2002). Also, we defined the Balmer continuum normalization in terms of the continuum luminosity at 3000 Å, since we expect the former to be a fraction of the latter, and this parametrization allows us to always start with a reasonable guess parameter ($A = 0.1$).

¹⁰ <http://www.physics.wisc.edu/craigm/idl/fitting.html>

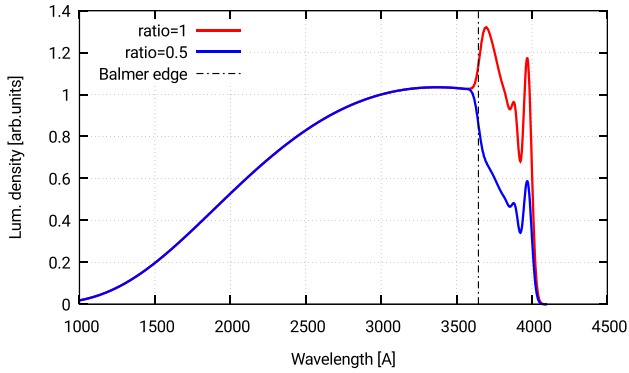


Figure 1. Plot of the Balmer continuum (at wavelengths shorter than the edge, 3645 Å) and pseudo-continuum template (at wavelengths longer than the edge). The latter is the superposition of high-order Balmer lines ($7 \leq n \leq 50$, i.e. H ϵ and higher). The ratio R of the pseudo-continuum to continuum luminosities at the Balmer edge is a free parameter in the fit. In the above plot, two ratios are shown: 1 (red) and 0.5 (blue). The two templates overlaps at wavelengths shorter than the edge, hence only the blue is visible.

The QSFIT Balmer component also accounts for high-order Balmer lines ($7 \leq n \leq 50$, i.e. H ϵ and higher) which are not fitted individually (Section 2.6) and blends into a Balmer ‘pseudo-continuum’ at wavelengths longer than 3645 Å. The line ratios are taken from Storey & Hummer (1995), for a fixed electron density of 10^9 cm^{-3} and a temperature of 15 000 K. The luminosity ratio R of the high-order blended line complex to the Balmer continuum (at the Balmer edge) is a free parameter in the fit.

The whole component (sum of Balmer continuum and high-order lines) is finally broadened by a Gaussian profile of 5000 km s^{-1} . A plot of the Balmer template, for two values of the ratio of the high-order blended line complex to the Balmer continuum (1 and 0.5, respectively) is shown in Fig. 1. The electron temperature and density, optical depth at Balmer edge and width of broadening profiles are fixed by default, but can be left free to vary for specific purposes. Also, for sources with $z \geq 1.1$, we observe the spectrum at wavelengths shorter than $\sim 4000 \text{ Å}$, i.e. we miss the most characterizing part of the Balmer continuum. At $\lambda \leq 3500 \text{ Å}$, the Balmer continuum resembles a simple power law, and it would become degenerate with the continuum component. Therefore, we fixed $A = 0.1$ and $R = 0.3$ for sources with $z \geq 1.1$ (see Fig. 23).

2.4 Host galaxy

The host galaxy may contribute to the observed luminosity of an AGN, especially at optical/IR wavelengths. It is therefore necessary to disentangle its contribution from the AGN continuum to provide reliable estimates of the latter (e.g. Vestergaard et al. 2008). The galaxy spectra typically show a change of slope at 4000 Å , corresponding to the Wien’s tail of blackbody spectra from stars, as shown in Fig. 2. This change of slope can be detected in SDSS spectra of low-luminosity, low-redshift ($z \leq 0.8$) AGN (e.g. fig. 3 in Calderone, Sbarrato & Ghisellini 2012) and can be used in automatic fitting procedures to estimate the host galaxy contribution.

In the current QSFIT implementation, we used a simulated 5 Gyr old elliptical galaxy template (Silva et al. 1998; Polletta et al. 2007) with a normalization factor as the only parameter. The template is normalized to have luminosity density of 1 at 5500 Å , hence the value of the normalization parameter is the luminosity density of the

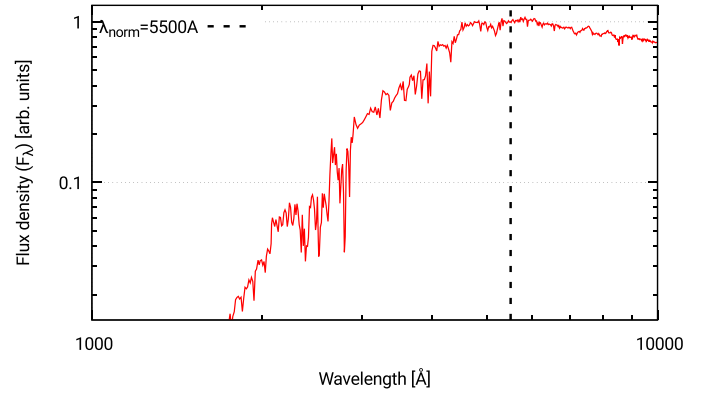


Figure 2. The normalized elliptical galaxy template from Polletta et al. (2007) used to model the host galaxy emission.

host galaxy at 5500 Å . We chose the elliptical template since luminous AGN (as those considered in the QSFIT catalogue, Section 3) are likely hosted in elliptical galaxies (e.g. Floyd et al. 2004). For sources with $z > 0.8$, we disabled the host galaxy component, since the SDSS rest-frame spectral range covers wavelengths shorter than 5000 Å , providing weak constraints on the host galaxy component.

Our approach in fitting the host galaxy contribution with a single template is clearly simplistic, but it allowed us to homogeneously analyse a very large sample. More sophisticated AGN/host galaxy decomposition methods have been proposed by other authors (e.g. Barth et al. 2015; Matsuoka et al. 2015), and could in principle be used also with QSFIT by customizing it for specific needs (Appendix B).

2.5 Blended iron lines

Modelling of both broad and narrow (permitted and forbidden) iron lines is essential for a proper estimate of broad-band continuum and contaminated lines (such as Mg II and H β), especially if the source shows small widths of broad lines (Boroson & Green 1992). In the current QSFIT implementation, we used the iron template of Vestergaard & Wilkes (2001) at UV wavelengths, and the iron template of Véron-Cetty et al. (2004) at optical wavelengths.

2.5.1 Iron emission lines at optical wavelengths

At optical wavelengths, we used the spectral decomposition of I Zw 1 made by Véron-Cetty et al. (2004), spanning the wavelength range from 3500 to 7200 Å (rest frame).

To compare the observed data with the optical iron template, we prepared a grid of templates, each with its own value of FWHM. To generate each template we sum a Gaussian profile (of the given FWHM) for each line listed in tables A.1 and A.2 of Véron-Cetty et al. (2004, excluding the hydrogen Balmer lines). The relative line intensities are fixed to those reported in the tables,¹¹ and the final template is normalized to 1. Hence, the model parameters are the width of the lines (which identifies a template in our grid) and its overall normalization. The ‘broad’ and ‘narrow’ templates (tables A.1 and A.2, respectively) are kept separated, each with its own parameters. The FWHM in our grid vary from 10^3 to 10^4 km s^{-1} for the broad templates, and from 10^2 to 10^3 km s^{-1} for the narrow

¹¹ We considered only the William Herschel Telescope intensities.

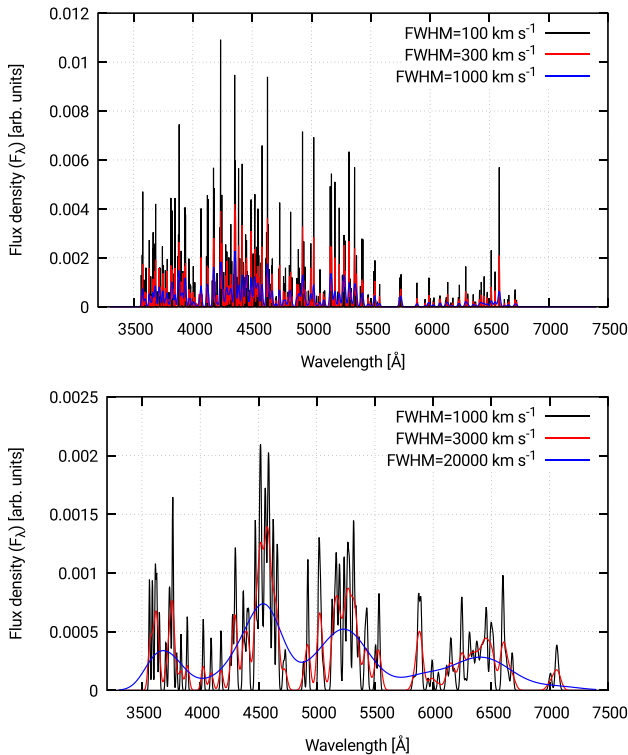


Figure 3. Three examples of narrow (upper panel) and broad (lower panel) optical iron templates used in the fit. These templates were obtained by summing a Gaussian line profile for each entry in tables A.1 and A.2 of Véron-Cetty et al. (2004) (excluding the hydrogen Balmer lines). The FWHM of the template is the FWHM of the Gaussian profile used to generate the template. All these templates are normalized to 1.

templates. A plot of the iron template, for three values of FWHM (10^3 , 3×10^3 and 10^4 km s $^{-1}$) is shown in Fig. 3. The iron emission at optical wavelengths is typically weak, and the SDSS spectral resolution is not sufficient to provide robust constraints to the FWHM values, hence we keep it fixed at 3000 km s $^{-1}$ for the broad component and to 500 km s $^{-1}$ for the narrow component. These constraints can be relaxed when using data with higher spectral resolution.

2.5.2 Iron emission lines at UV wavelengths

We considered the ‘B’ template and the Fe II UV191 and Fe III UV47 multiplets of the iron template described in Vestergaard & Wilkes (2001), spanning the wavelength range from 1250 to 3090 Å. The original template is based on the observations of I Zw 1, whose FWHM of broad lines is 900 km s $^{-1}$. To adapt the template to sources with broader emission lines, we prepared a grid of broadened templates by convolving the original one with a set of Gaussian profiles, whose FWHM range from 10^3 to 2×10^4 km s $^{-1}$. In this case, there is no separation between ‘broad’ and ‘narrow’ contributions, since the original template in Vestergaard & Wilkes (2001) is the sum of both. A plot of the template, for three values of FWHM (10^3 , 3×10^3 and 2×10^4 km s $^{-1}$) is shown in Fig. 4.

The model parameters are the equivalent width (integrated over the whole wavelength range 1250 to 3090 Å) and the FWHM of the profile used to broaden the template. The FWHM values are constrained in the range 10^3 – 10^4 km s $^{-1}$. The SDSS spectral resolution is not sufficient to provide robust constraints to the FWHM

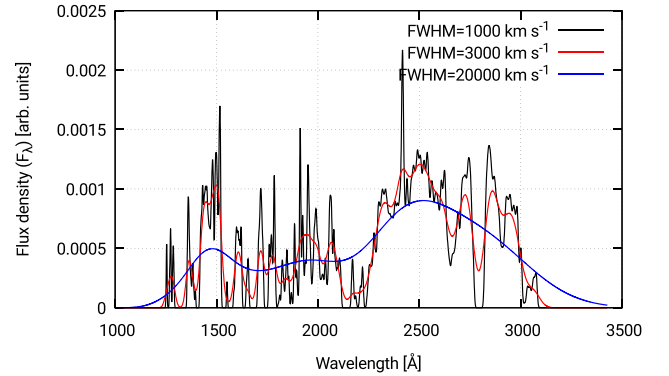


Figure 4. Three examples of the broadened and normalized UV iron templates. These templates were obtained by convolving the original iron template in Vestergaard & Wilkes (2001) with a Gaussian of FWHM equal to 10^3 , 3×10^3 and 2×10^4 km s $^{-1}$ respectively.

value, hence we keep it fixed at 3000 km s $^{-1}$. This constraint can be relaxed when using data with higher spectral resolution. Also, for sources with $z \leq 0.4$, we fix the iron UV equivalent width to 138 Å (i.e. the median of the values measured at larger redshifts), since only a small part of the whole template is visible, and the iron normalization would become degenerate with the continuum slope. Moreover fixing the UV iron helps in constraining the continuum slope to have similar distributions at redshifts below and above 0.4.

2.6 Main emission lines

The broad and narrow emission lines considered in QSFIT are shown in Table 1. Each line is modelled with a Gaussian profile, whose fitting parameters are the total (integrated) line luminosity, the FWHM and velocity offset (with respect to the reference wavelength) of the line profile. The FWHM of narrow and broad lines is constrained in the range $[10^2, 2 \times 10^3]$ and $[9 \times 10^2, 1.5 \times 10^4]$ km s $^{-1}$, respectively. The velocity offset of narrow and broad lines is constrained in the range $\pm 10^3$ and $\pm 3 \times 10^3$ km s $^{-1}$, respectively. For the H α and H β emission lines, we used both broad and narrow components, but we had to limit the FWHM of the narrow component to $[10^2, 10^3]$ km s $^{-1}$ to allow a good decomposition of the line profile. Also, we had to limit the Mg II broad component velocity offset to $\pm 10^3$ km s $^{-1}$ to avoid degeneracies with the UV iron template. Finally, we used three components to fit the H α to account for the very broad feature below that line. The FWHM of the H α ‘base’ line is constrained in the range $[10^4, 3 \times 10^4]$ km s $^{-1}$.

A line component is ignored if the resulting line profile would fall outside (even partially) the available wavelength range. Moreover, a line component is ignored if the amount of dropped spectral channels (whose quality mask is not 0, Section 2.8.1), in the wavelengths relevant to the emission line, exceeds 40 per cent. These precautions were taken to ensure the emission-line parameters can actually be constrained by the available data.

2.7 ‘Unknown’ emission lines

Beyond the ‘known’ emission lines (Section 2.6) expected to be relevant in any AGN, we also considered a list of 10 ‘unknown’ emission lines, i.e. not a priori associated with any known line. These emission lines are added after all the other components (see Section 2.8), and their initial wavelength is set at the position where

there is the maximum positive residual. By considering these further components, we may account for the lack of an iron template in the wavelength range 3100–3500 Å, or for asymmetric profiles in known emission lines.

2.8 Fitting procedure and data reduction

In this section, we describe the spectral analysis procedure, which runs through four distinct steps: (i) preparation of the spectrum, (ii) model fitting and (iii) data reduction.

2.8.1 Spectrum preparation

The current QSFIT implementation operates only on spectra observed by the SDSS-DR10 (York et al. 2000; Ahn et al. 2014).

To prepare the spectrum, we drop the 100 spectral channels at the beginning and end of each spectrum to avoid artefacts from instrument or pipeline. Then, we drop the spectral channels whose quality mask is not equal to zero. If the remaining ‘good’ channels accounts for less than 75 per cent of the original channels, corresponding to ~ 2700 channels, QSFIT issues an error and skips the analysis. In the vast majority of cases, the final wavelength range in the observer frame is ~ 4150 – 8740 Å, and the number of channels is ~ 3450 . The spectral resolution is $\sim 150 \text{ km s}^{-1}$.

The spectrum is corrected for Galactic extinction using the CCM parametrization by Cardelli, Clayton & Mathis (1989) and O’Donnell (1994), assuming a total selective extinction $A(V)/E(B - V) = 3.1$. We neglected any intrinsic (i.e. rest frame) reddening of the spectrum, since most of sources show small or no reddening at all (e.g. Grupe et al. 2010). To build the QSFIT catalogue (Section 3), we used the same colour excess and redshift estimates as S11. Finally, the spectrum is transformed to the rest frame using the redshift estimate. The fitting process is performed by comparing the data and the model, in units of $10^{42} \text{ erg s}^{-1} \text{ Å}^{-1}$.

2.8.2 Model fitting

The fitting process occurs by varying the component parameters until the χ^2 , as obtained by comparing the data and the model, is minimized. The minimization procedure follows a Levenberg–Marquardt algorithm, hence the result may depend significantly from the initial (guess) parameter values. In order to develop an ‘automatic’ spectral analysis (without human intervention), we should find a proper way to get rid of such dependency on initial parameters. We found that a promising approach is to build the model step by step, i.e. by iteratively adding a component and re-running the minimization procedure.

In the following, we will discuss the five steps involved in the fitting process. The plots of data, model and residuals at each step of the fitting procedure are shown in Fig. 5. In order to produce clearer plots all data, as well as the model, were rebinned by a factor of 5. The analysis, however, has been carried out at full resolution. Additionally, the catalogue browser web tools allows the user to view the data using any rebinning factor.

(i) The first components being added are the AGN continuum (actually a power law, Section 2.2) and the host galaxy template (Section 2.4). A first minimization is performed, which will bring the ‘best-fitting’ model to pass through the data, without accounting for any emission line. The amount of positive residuals, calculated as $(\text{data} - \text{model})/\text{uncertainty}$, is approximately equal to the amount of negative residual.

(ii) In order to provide room for further components (namely the emission lines), we lower the continuum normalization until the positive residuals reach ~ 90 per cent,¹² and fix all parameters of both AGN continuum and host galaxy components for the next iterations.

(iii) We add the components for the iron templates at UV and optical wavelength (Section 2.5). The optical template is enabled for sources with $z \lesssim 1$, while the UV iron template is always enabled, although the equivalent width is fixed to 138 Å if $z < 0.4$ (Section 2.5.2). Then, we re-run the fitting process by varying just the iron template parameters and we fix them at their best-fitting values for the next iterations.

(iv) Now we add the ‘known’ emission lines, and ensure there are enough spectral channels to constrain the line parameters, otherwise the line is ignored (Section 2.8.1). After re-running the minimization, we fix all the emission-line parameters for the next iterations.

(v) we iteratively add the ‘unknown’ emission lines (Section 2.7), i.e. lines which are not a priori associated with any specific transition. The centre wavelength of these lines is first set at the position of the maximum positive residual, i.e. where most likely there is some further emission component, and then left free to change during the fitting process. We repeat this procedure up to 10 times. Finally, we re-run the fitting process with all parameters, for all components, left free to vary. At the end, we check which ‘unknown’ line component is actually required to provide a better fit: if the line luminosity uncertainty is greater than 3 times the luminosity we disable the component and re-run the fit.

2.8.3 Data reduction

After the fitting process, we use the resulting best-fitting parameter values to compile a list of spectral properties for the source being analysed. In few cases, a further *reduction* is required to provide the final quantities. The details of the reduction process are as follows:

(i) the AGN continuum and galaxy luminosities are estimated directly on the continuum component values, in at least six wavelengths, five of them equally spaced in the (logarithmic) wavelength range $[\lambda_{\min} + 5 \text{ per cent}, \lambda_{\max} - 5 \text{ per cent}]$. We restricted the range by 5 per cent on both sides to avoid artefacts at the edges of the available spectrum. The other estimates are given at the fixed rest-frame wavelengths 1450, 2245, 3000, 4210 or 5100 Å, depending on the source redshift;

(ii) the ‘unknown’ lines are associated with the ‘known’ emission lines if the centre wavelength of the latter falls within the FWHM of the former. The association is performed only if the ‘unknown’ line has an FWHM of at least 1000 km s^{-1} and the ‘known’ line being considered is a broad line (narrow lines are always modelled with a single component). If two or more emission-line components are associated, then we compute the total line profile as the sum of individual line profiles, and estimate the relevant quantities as follows:

(a) the line luminosity is the sum of integrated luminosities of each component;

(b) the FWHM is the full width of the total profile at half the maximum of the total profile peak;

¹² Although there appear to be some arbitrariness in performing this step, we found that the final results (after all the remaining steps have been performed) depends weakly on the fraction of positive residuals, as long as the latter is above ~ 70 per cent.

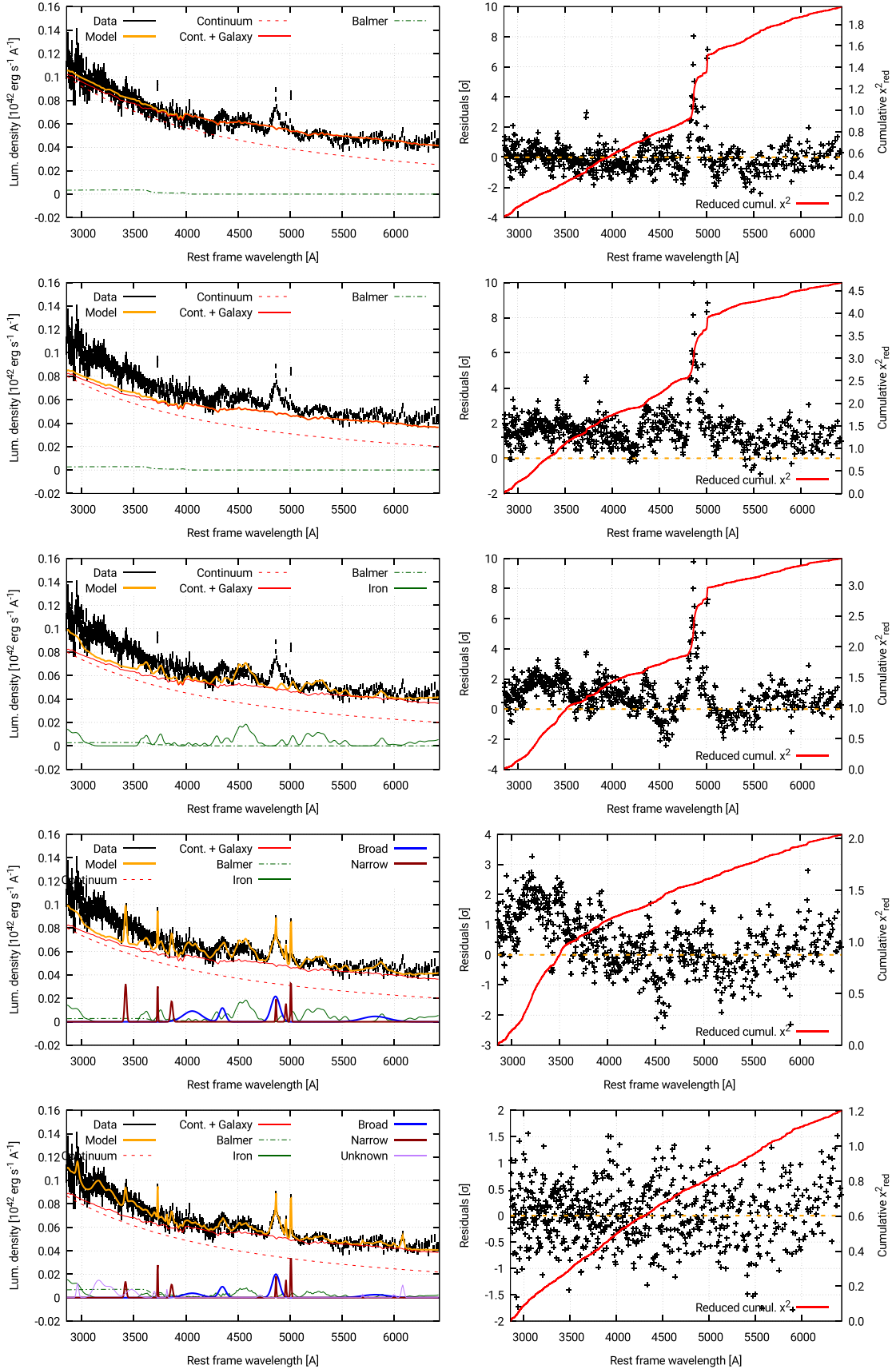


Figure 5. The plots in the left-hand column show the comparison between the data and the overall model, as the individual components are being added. The five steps of the fitting procedure (corresponding to the five rows in the plots above) are described in Section 2.8.2. In order to produce clearer plots, all data, as well as the model, were rebinned by a factor of 5. The final plot at full resolution is shown in Fig. 6.

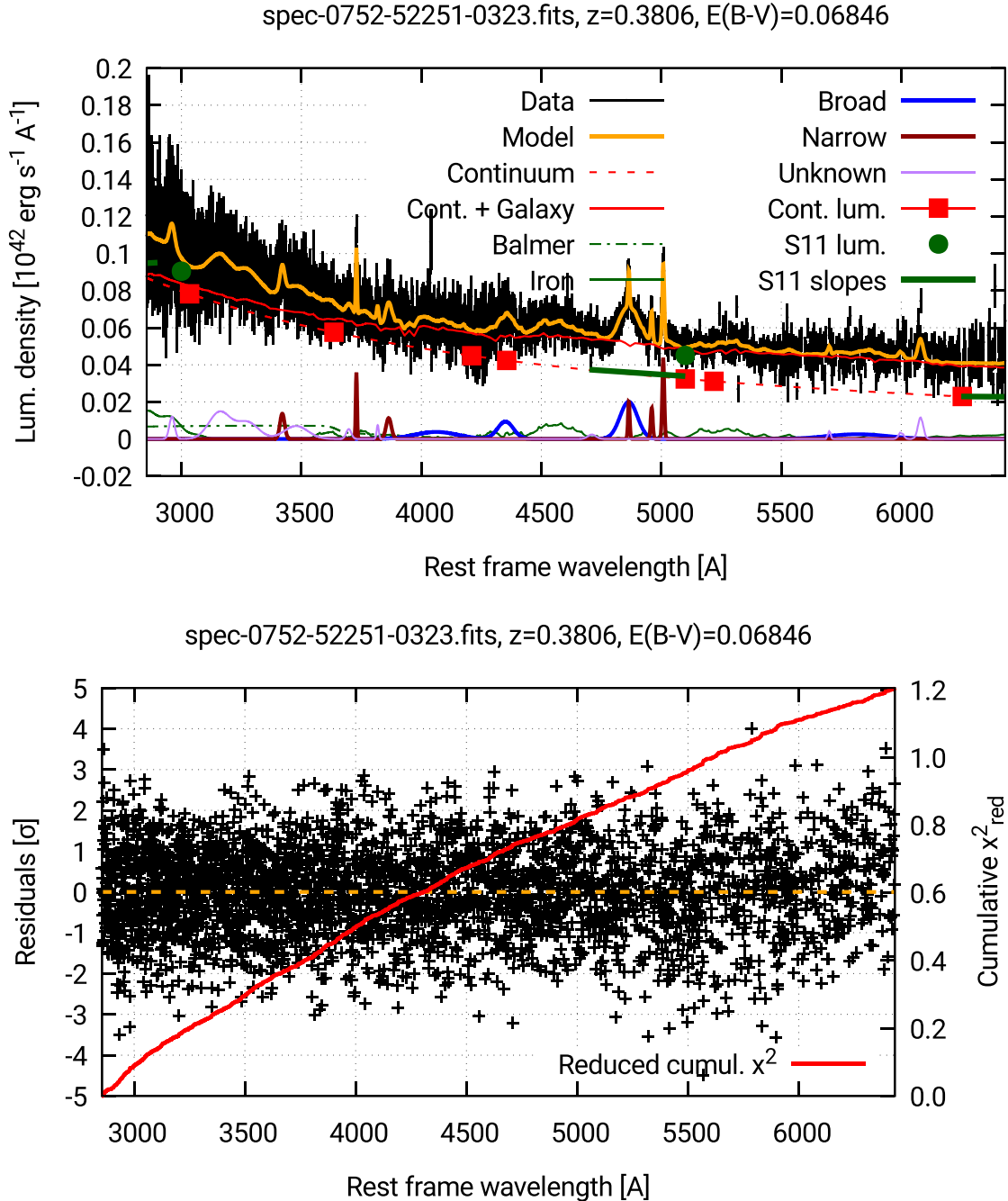


Figure 6. Comparison of the QSFIT model and SDSS data: the upper panel shows the comparison between the SDSS data with their error bars (black vertical lines) and the QSFIT model (orange line); the lower panel shows the residuals (data – model) in units of 1σ uncertainties in the data (black cross symbol) and the cumulative χ^2_{red} (red line, values on the right-hand axis). The upper panel also shows the individual components in the QSFIT model: the dashed red line is the continuum component (Section 2.2), while the solid red line is the sum of continuum and host galaxy components (Section 2.4); the dot-dashed green line is the Balmer component (Section 2.3); the solid green lines are the optical (Section 2.5.1) and UV (Section 2.5.2) iron templates; the sum of all broad and of all narrow emission-line (Section 2.6) components are shown with blue and brown lines, respectively. The sum of all ‘unknown’ lines (Section 2.7) is shown with a purple solid line. The red square and green circle symbols are the continuum luminosity estimated by QSFIT and S11, respectively. The thick green line segments show the slopes estimated by S11 (anchored on the short-wavelength side on QSFIT continuum component).

(c) the velocity offset is calculated by considering the offset between the total profile peak position and the emission-line reference wavelength.

The uncertainties on the luminosity is computed as the weighted sum of the luminosity uncertainty on each component. The uncer-

tainties on FWHM and velocity offset are computed as the weighted average of the corresponding uncertainties in the fit parameters, with the weight given by the line intensities. These parameters are typically degenerate hence the uncertainties may be underestimated, see Appendix B4 for a different approach to estimate parameter uncertainties;

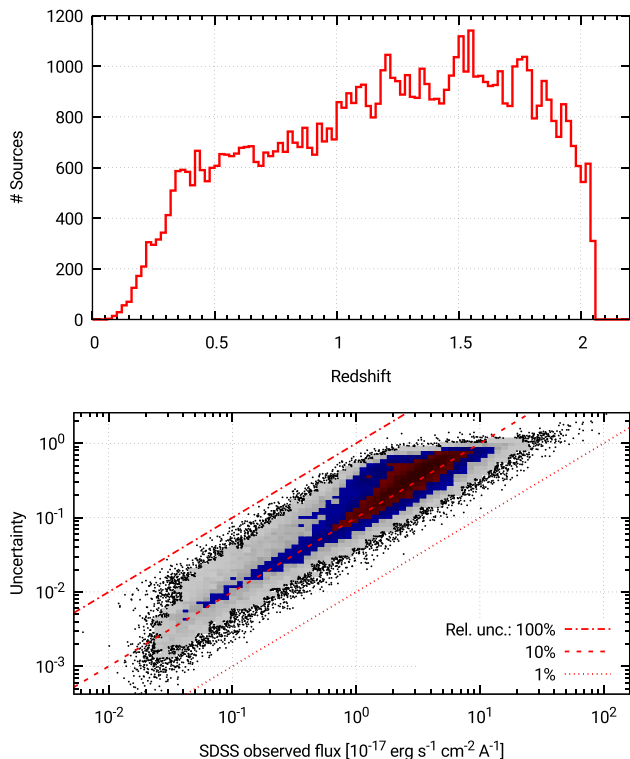


Figure 7. Distribution of redshift (upper panel) and SDSS observed flux density versus its uncertainty (lower panel). See Section 3 for a description of the colours used in the 2D histogram. The dotted, dashed and dot-dashed red lines are drawn in the region of points whose relative uncertainty is 1 per cent, 10 per cent and 100 per cent, respectively.

(iii) beyond the spectral quantities, we also provide several ‘quality flags’ to assess the reliability of the automatic results. If a flag is raised for a given source then the corresponding quantity should be used cautiously, and a visual analysis of the spectrum is recommended. To produce the plots in the following sections, we considered only the sources with no quality flag raised (i.e. whose quality flag is 0). The list of all quality flags and their meaning is discussed in Appendix A, while the tables in Appendix C summarize the fractions of sources which raised a quality flag.

All quantities calculated by QSFIT are described in Appendix A.

The uncertainties provided by QSFIT are calculated using the Fisher matrix method (e.g. Heavens 2009; Andrae 2010), by assuming Gaussian uncertainties in the SDSS data, negligible correlation among model parameters and symmetric uncertainty intervals. These assumptions are not always justifiable, hence QSFIT uncertainties should be considered as rough estimates of parameter uncertainties. A better approach, although significantly more demanding in terms of computational time, is to use the Monte Carlo resampling method as discussed in Appendix B4. In the following, we will always consider the Fisher matrix uncertainties provided by QSFIT.

2.9 QSFIT plots

After analysing the data QSFIT produces two plots using `GNUPLOT`.¹³ An example is shown in Fig. 6: the upper panel shows the com-

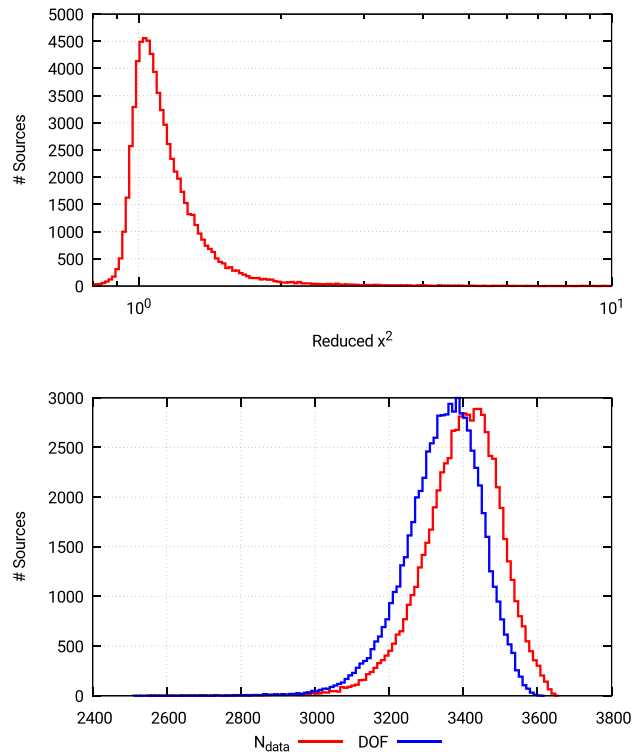


Figure 8. Upper panel: distribution of reduced χ^2_{red} in the QSFIT catalogue. Lower panel: distribution of number of spectral channels in the SDSS data (N_{data} , red line) and of DOF in the fit (blue line).

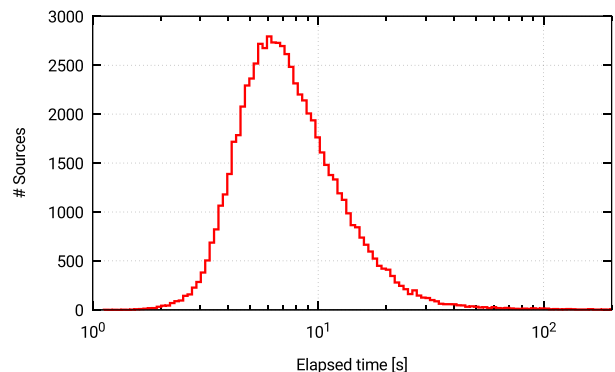


Figure 9. Distribution of elapsed time (in seconds) to perform a QSFIT analysis on a modern laptop.

parison between the SDSS data with their error bars (black vertical lines) and the QSFIT model (orange line); the lower panel shows the residuals (data minus model) in unit of 1σ uncertainties in the data (black cross symbol) and the cumulative χ^2_{red} (red line, values on the right-hand axis). The upper panel also shows the individual component in the QSFIT model: the dashed red line is the continuum component (Section 2.2), while the solid red line is the sum of continuum and host galaxy components (Section 2.4); the dot-dashed green line is the Balmer component (Section 2.3); the solid green lines are the optical (Section 2.5.1) and UV (Section 2.5.2) iron templates; the sum of all broad and of all narrow emission-line (Section 2.6) components are shown with blue and brown lines, respectively. The sum of all ‘unknown’ lines (Section 2.7) is shown

¹³ <http://gnuplot.info/>

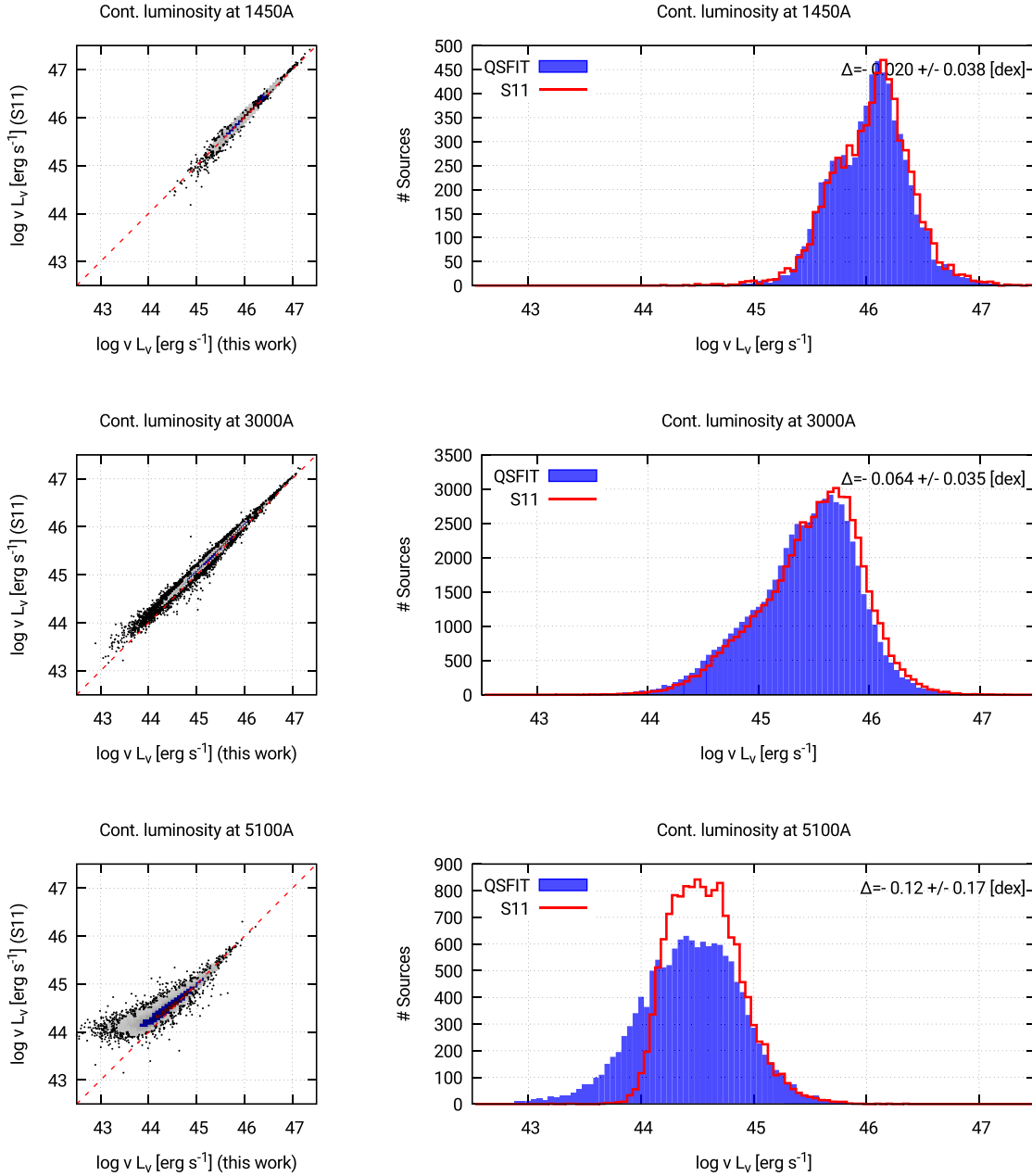


Figure 10. Comparison of QSFIT and S11 continuum luminosity estimates at 1350, 3000 and 5100 Å. The left-hand panels show the 2D histogram plots (see Section 3 for a description of such plots) with QSFIT values on the X-axis and S11 values on the Y-axis (the red dashed line is the equality line). In the right-hand panels, we compare the distribution of the two populations. The average and standard deviation of the difference between the two catalogues is shown in the upper right corner. The estimates are compatible in the two catalogues with a scatter of $\lesssim 0.04$ dex (~ 10 per cent) over ~ 2 dex, except for the 5100 Å case of low-luminosity AGN ($L_{5100} \lesssim 10^{44.5}$ erg s $^{-1}$), since the host galaxy contribution has been considered in QSFIT, but neglected in S11 (see also Section 5.2).

with a purple solid line. The red squared and green circle symbols are the continuum luminosity estimated by QSFIT and S11, respectively. The thick green line segments show the slopes estimated by S11 (anchored on the short-wavelength side on QSFIT continuum component).

In the example shown in Fig. 6, the comparison between the data and the QSFIT model yield a χ^2_{red} value of 1.19, with 3248 degrees of freedom (DOF). Given the very high number of DOF, the probability of obtaining a worst χ^2_{red} by random fluctuations is very low ($\ll 10^{-5}$), hence the model should actually be rejected on a statistical basis. Nevertheless, a qualitative inspection shows

that the residuals are equally distributed on positive and negative values, and the cumulative χ^2_{red} line (red line in lower panel) do not shows significant ‘jumps’ or discontinuities. Also, the spectral decomposition of the H β and [O III] region, as well as the iron emission, appear reasonable. Finally, the correlation among parameters (Appendix B4) implies that the actual number of DOF is larger than the quantity *spectrum channel* – *free parameters*, hence the quoted χ^2 is actually an upper limit for the actual value. In summary, we consider our spectral quantities reliable, even if the model should be rejected according to the χ^2_{red} goodness-of-fit test.

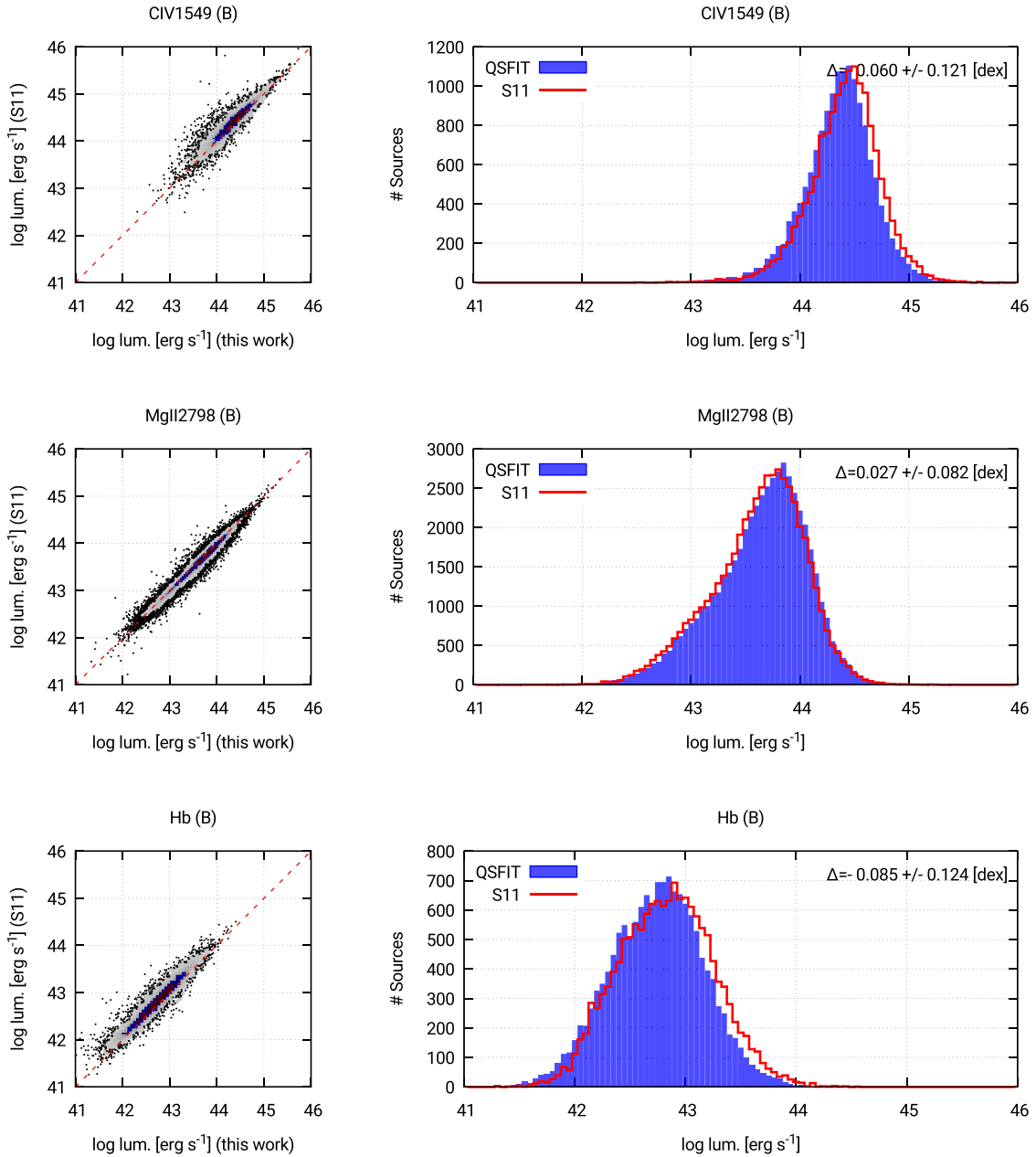


Figure 11. Comparison of QSFIT and S11 broad component luminosities of C iv 1549 Å, Mg ii 2798 Å and H β 4861 Å emission lines. The plot meaning is the same as in Fig. 10. The most prominent differences are in the C iv 1549 Å and H β 4861 Å emission lines, mainly due to the different continuum subtraction methods (Section 2.8).

3 THE QSFIT CATALOGUE

We used QSFIT to analyse a subsample of the fifth edition of the SDSS quasar catalogue (Schneider et al. 2010), i.e. the same sample analysed by Shen et al. (2011, the S11 catalogue). We selected all the objects in the S11 catalogue whose redshift is < 2.05 and whose BAL_FLAG is set to 0, i.e. they are not broad absorption line (BAL) quasars. This amounts to 79 986 sources. Then, we dropped all the sources whose number of ‘good’ spectral channels (quality mask = 0) was less than 75 per cent (Section 2.8). The final sample comprises 71 251 sources, which make up the QSFIT catalogue (version 1.2) of spectral quantities. Fig. 7 shows the

redshift distribution in our sample (upper panel) and the distribution of the SDSS flux density versus its uncertainty (lower panel).

The colours in the 2D histogram in the lower panel of Fig. 7 represent the count densities. The bins drawn with red and blue palettes contain, respectively, 40 per cent and 70 per cent of the sources, the grey, red and blue ones (taken together) contain 95 per cent of the sources. The remaining 5 per cent are shown with individual dots. This colour code is used for all the 2D plots in this paper.

For each source in our subsample, we performed the spectral analysis using the QSFIT procedure and collected all the results in the QSFIT catalogue. A similar work has already

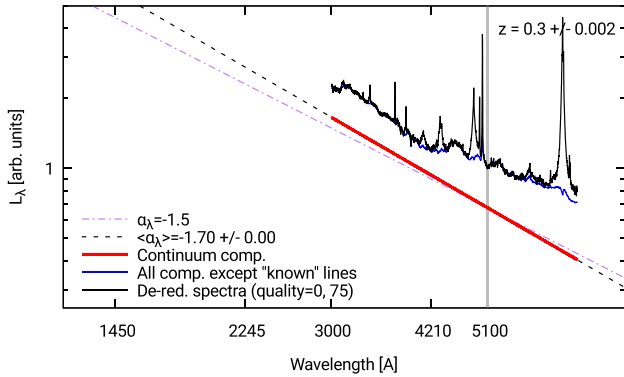


Figure 12. Comparison of the averaged spectrum of SDSS de-reddened spectra in the redshift bin $z = 0.3 \pm 0.002$ (black line) with the average spectrum of the sum of all QSFIT components, except the ‘known’ emission lines (blue line). The agreement between the blue and black lines is remarkably good over the entire wavelength range (except the regions of the main emission line). Hence, the blue line represent a reliable estimate of the continuum below the line (at least on average).

been done by Shen et al. (2011), although with a different approach (Section 4). In order to perform a meaningful comparison of our results and those from S11 (Section 4), we used exactly the same redshift and $E(B - V)$ values used by S11, as well as their cosmology ($H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$, $\Omega_M = 0.3$ and $\Omega_\Lambda = 0.7$).

In order to estimate the reliability of QSFIT results we performed a qualitative visual inspection of many of the QSFIT plots (Section 2.9) and found that our model usually provides a rather good representation of the data. A quantitative assessment of the agreement is given by the χ^2_{red} values, whose median value is 1.09 and is almost always smaller than 2 (Fig. 8, upper panel). The number of DOF is typically in the range [3100, 3600] with a median value of 3380 (Fig. 8, lower panel). Despite the rather low values of χ^2_{red} , the goodness-of-fit tests typically fail, since the number of DOF is very high. Hence, the QSFIT model is too simple to account for all the details available in the data, and further components are likely required (e.g. emission lines, improved iron and host galaxy templates, etc.). Nevertheless, as discussed in Section 2.9, a qualitative examination of the QSFIT plots, as well as a comparison with the S11 catalogue (Section 4), typically ensure that the spectral quantities can be considered reliable estimates. The ‘average’ reliability of the QSFIT estimates can also be assessed by comparing the average SDSS spectra with the average QSFIT components (see Fig. 12 and the discussion in Section 5.2).

Moreover, the current QSFIT implementation aims to provide a reasonably good fit of SDSS data in the vast majority of the cases. In order to provide a more accurate representation of the data, the QSFIT procedure can be customized and tuned for the analysis of a specific source (Appendix B).

The QSFIT catalogue is available as a FITS file,¹⁴ where we report all the columns discussed in Section 2.8.3, as well as all the columns from the S11 catalogue. The QSFIT catalogue is also browsable through a web tool¹⁵ which allows the user to search for a specific source, sort sources according to a parameter, e.g. redshift, interac-

tively browse the spectra inspecting the various model components, eventually rebinning the spectra, and finally download the plots (comparison of data and model, fit residuals) and the QSFIT log file for a specific source. The rather low computational time (~ 7 s on a modern day laptop, Fig. 9) required to analyse a source enables to re-run the analysis of individual source, re-build the whole QSFIT catalogue, expand it, etc., without the need for expensive computational resources.

4 COMPARISON OF S11 AND QSFIT CATALOGUES

In this section, we will compare our results with those from Shen et al. (2011, S11) in order to assess the reliability of QSFIT results. Although the purpose of both the S11 and QSFIT catalogues is similar, the adopted analysis algorithms show significant differences. The most relevant are:

(i) S11 evaluates the continuum underlying the emission lines over a rather narrow wavelength range, in the neighbourhood of the line, whereas QSFIT uses a continuum component spanning the whole available wavelength range (Section 2.2);

(ii) S11 fits the continuum and iron contributions first, then subtract them from the observed flux, and finally fits the emission lines. QSFIT follows a similar approach, but at the last step all parameters (continuum + galaxy + iron + emission lines) are free to vary simultaneously;

(iii) S11 neglects the host galaxy contamination, while QSFIT uses an elliptical galaxy template to account for such contribution (Section 2.4);

(iv) S11 uses the iron template from Boroson & Green (1992) for the $H\alpha$ and $H\beta$ emission lines, while QSFIT uses the template from Véron-Cetty et al. (2004);

(v) S11 models the broad emission lines with either one, two or three Gaussian profiles, according to which case provide the best fit. Also, the $[\text{O III}]4959 \text{ \AA}$ and $[\text{O III}]5007 \text{ \AA}$ are both modelled with two Gaussian profiles. QSFIT always starts with a single Gaussian profile for each broad and narrow lines. If an ‘unknown’ emission line (Section 2.7) lies sufficiently close to a ‘known’ line, it can be associated with the latter (Section 2.8.2) and the final profile may be modelled with more than one Gaussian profile. Table C4 shows how many lines were fitted with more than one component.

Despite the differences in the adopted analysis algorithms, in several cases the spectral estimates in the S11 and QSFIT catalogues show a very good correlation, as is the case for, e.g. the continuum luminosity estimates and the $\text{C IV } 1549 \text{ \AA}$, $\text{Mg II } 2798 \text{ \AA}$ and $H\beta$ $4861 \text{ \AA}$ emission-line luminosities. Fig. 10 shows the comparison of QSFIT and S11 continuum luminosity estimates at 1350, 3000 and $5100 \text{ \AA}$. The left-hand panels show the 2D histogram plots (see Section 3 for a description of such plots) with QSFIT values on the X-axis and S11 values on the Y-axis (the red dashed line is the equality line). The right-hand panels show the comparisons between the population distributions. The average and standard deviation of the difference between the two catalogues is shown in the upper right corner. The estimates are compatible in the two catalogues with a scatter of $\lesssim 0.04 \text{ dex}$ (~ 10 per cent) over $\sim 2 \text{ dex}$, except for the $5100 \text{ \AA}$ case of low-luminosity AGN ($L_{5100} \lesssim 10^{44.5} \text{ erg s}^{-1}$), since the host galaxy contribution has been considered in QSFIT, but neglected in S11 (see also Section 5.2).

Fig. 11 shows the same comparisons for the broad component luminosities of $\text{C IV } 1549 \text{ \AA}$, $\text{Mg II } 2798 \text{ \AA}$ and $H\beta$ $4861 \text{ \AA}$ emission lines: also in this case the correlation is quite good, although the

¹⁴ http://qsfit.inaf.it/cat_1.2/fits/qsfit_1.2.fits

¹⁵ <http://qsfit.inaf.it/>

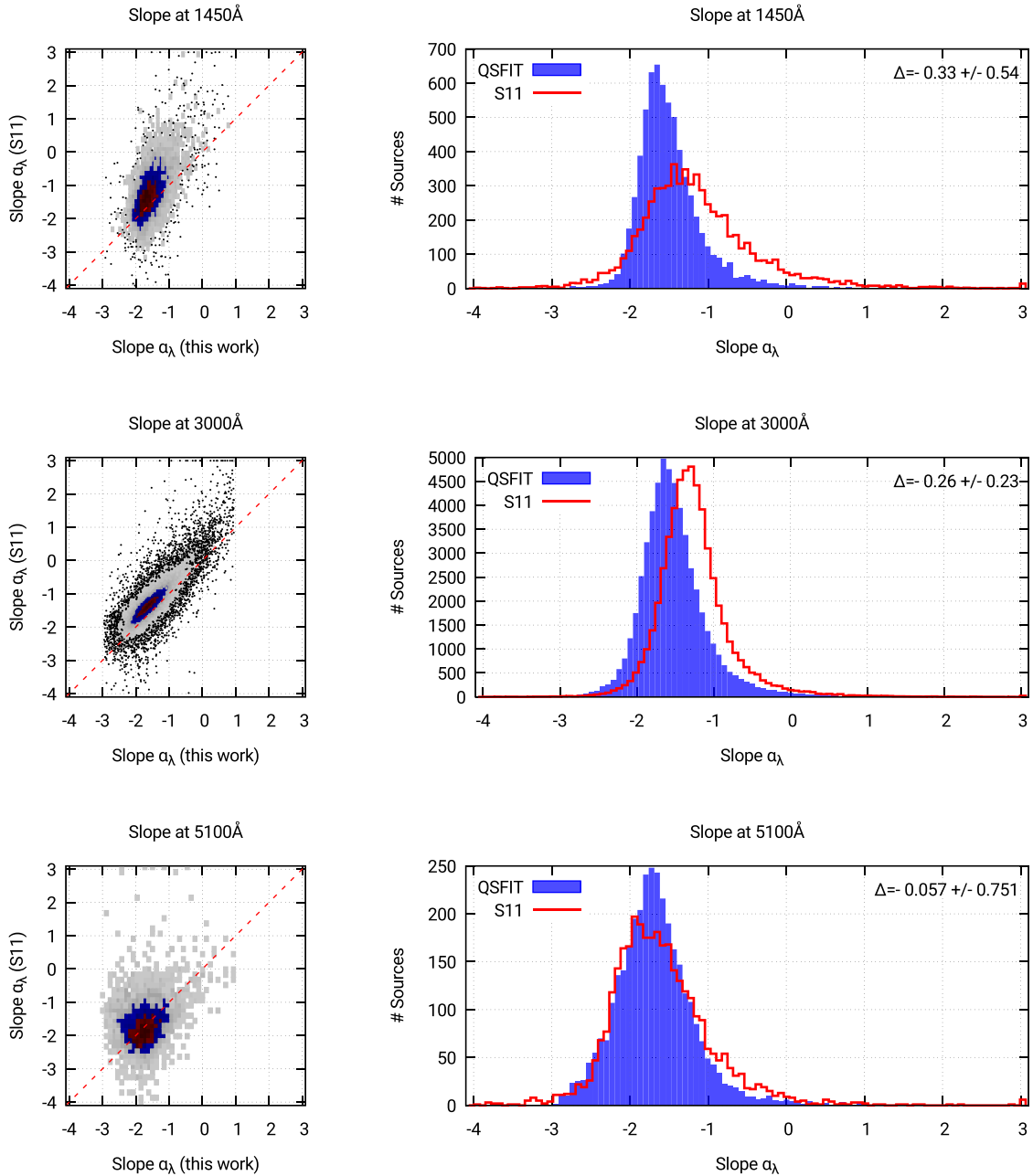


Figure 13. Comparison of our estimates of continuum slopes at 1350, 3000 and 5100 Å with those in S11. The estimates are not compatible because the continuum has been estimated using two different approaches: in this work we used a single power law extending through the entire observed wavelength range, while S11 used a power law in a narrow wavelength range.

scattering is larger ($\lesssim 0.18$ dex ~ 50 percent) over ~ 2 dex. The most prominent differences are in the C IV 1549 Å and H β 4861 Å emission lines, and are mainly due to the different continuum assumed below the lines (Section 2.8). In QSFIT such continuum is the sum of all the components except the ‘known’ emission lines. The reliability of such continuum can be inferred, at least on average, by comparing the averaged SDSS de-reddened spectra (in a narrow redshift bin) with the averaged sum of all QSFIT components except the ‘known’ emission lines. Such comparison is shown in Fig. 12, with the averaged spectrum of SDSS de-reddened spectra in the redshift bin $z = 0.3 \pm 0.002$ shown with a black line, and the average spectrum of the sum of all QSFIT components except

the ‘known’ emission lines shown with a blue line. The agreement between the blue and black lines is remarkably good over the entire wavelength range (except the regions of the main emission line). Hence, the blue line represents a reliable estimate of the continuum below the line (at least on average). The averaged (or composite) spectra are further discussed in Section 5.2.

The very good agreement in continuum and emission-line luminosities, spanning ~ 2 dex, stem from both an accurate spectral analysis, and the relatively large redshift range considered. Indeed, when considering quantities which do not depend directly on redshift the correlation is worse, although still present. This is the case for, e.g. the continuum slopes (Fig. 13) and the emission-line widths

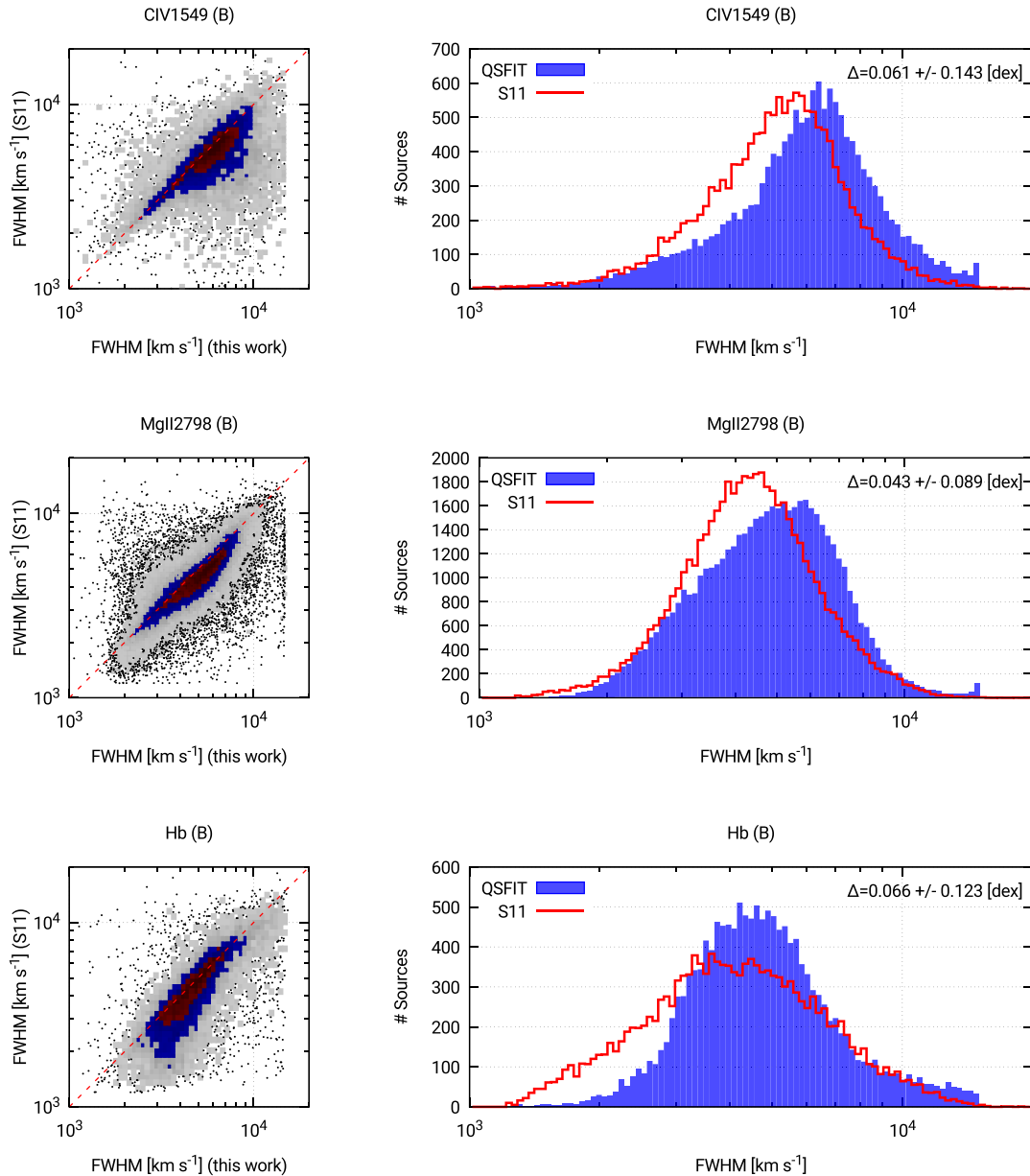


Figure 14. Comparison of QSFIT and S11 broad component widths of CIV 1549 Å, MgII 2798 Å and Hβ 4861 Å emission lines. The plot meaning is the same as in Fig. 10. The discrepancies are mainly due to the different line profile decomposition.

(Fig. 14). The discrepancies are mainly due to the different adopted algorithms. For instance, in case of continuum slopes, we refer to the broad-band continuum slopes, while S11 refer to the local continuum in the neighbourhood of an emission line.¹⁶ Moreover, the emission-line widths show a significant scattering ($\lesssim 0.16$ dex, ~ 45 per cent, over less than 1 dex), since the line profile decomposition algorithm is very different. These findings show that estimating the emission-line widths is not an easy and straightforward task, and the availability of a commonly accepted procedure would be very useful.

The rather low average differences between the populations of luminosities and line widths discussed above allow us to conclude

that, at least on average, the QSFIT estimates are as reliable as the S11 ones. Since there is not yet a complete understanding of the emission processes generating the broad-band continuum or the emission lines, there is not a clear way to prefer either the S11, QSFIT, or any other catalogue. The advantage in using the QSFIT one is simply that the analysis is performed in an automatic (i.e. fast on large data sets), replicable and shareable way. Moreover, it allows users to customize and fine tune the analysis for specific sources.

5 DISCUSSION

The main purpose of the QSFIT catalogue is to analyse the average spectral properties of Type 1 AGNs by taking advantages of the large SDSS data set, instead of focusing on individual sources

¹⁶ The plots available in the QSFIT webpage allow for a quick comparison of the spectral slopes of both the QSFIT and S11 catalogues.

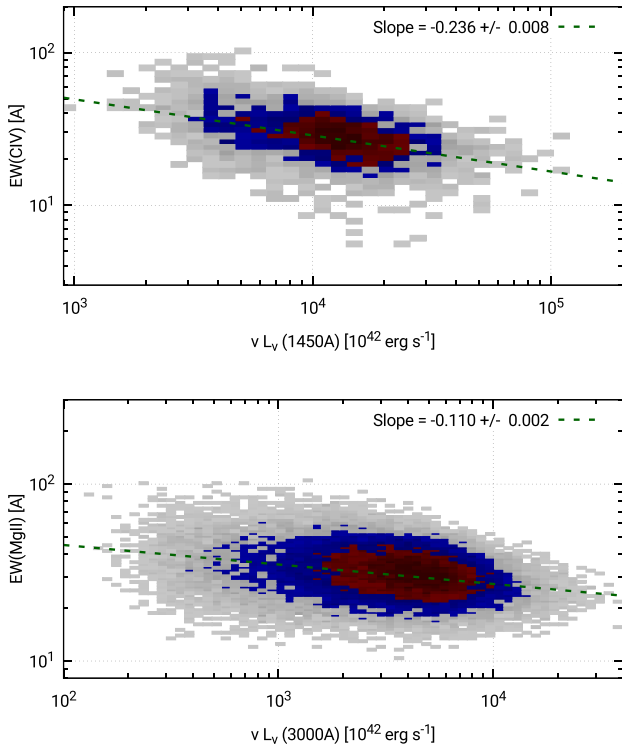


Figure 15. The Baldwin effect of the C IV 1549 Å (upper panel) and Mg II 2798 Å emission lines (lower panel) using the data from the QSFIT catalogue. The slopes are very similar to those reported in Green, Forster & Kuraszkiewicz (2001), although the sample used here is significantly larger.

whose spectrum may be affected by low signal-to-noise ratio, or be peculiar in some aspect. A thorough analysis of the QSFIT results is, however, beyond the purpose of this work, and will be discussed in forthcoming papers. Here, we will briefly explore the QSFIT data to check whether they follow the empirical relations already known in literature (Section 5.1), discuss the reliability of the QSFIT continuum slope estimates (Section 5.2) and present a few of the major results (Section 5.3).

5.1 Empirical relations

Several empirical relations have been identified among AGN spectral quantities (e.g. Sulentic, Marziani & Dultzin-Hacyan 2000). One of the most important ones is the ‘Baldwin effect’ (Baldwin 1977), i.e. the correlation between the C IV 1549 Å equivalent width and the continuum luminosity. As new data became available similar correlations have been found for other emission lines (e.g. Baldwin, Wampler & Gaskell 1989; Kinney, Rivolo & Koratkar 1990; Pogge & Peterson 1992; Green, Forster & Kuraszkiewicz 2001). In Fig. 15, we show the Baldwin effect for the C IV 1549 Å (upper panel) and Mg II 2798 Å emission lines (lower panel) using the data from the QSFIT catalogue. The slopes (shown in the upper right corner of the plots) are very similar to those reported in Green et al. (2001), -0.227 ± 0.025 and -0.187 ± 0.017 for C IV 1549 Å and Mg II 2798 Å, respectively, although the sample used here is significantly larger.

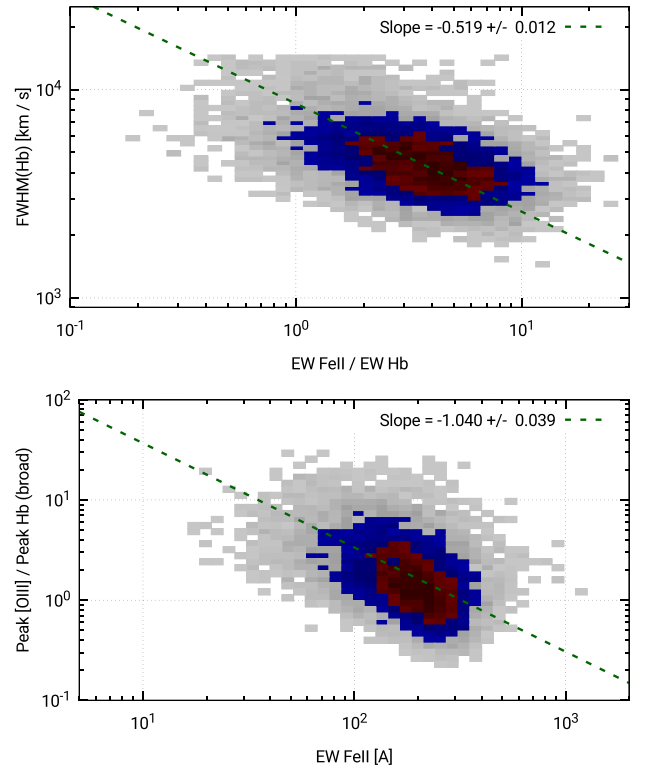


Figure 16. Correlations between the width of the broad H β emission line (upper panel) and the peak luminosity ratios between [O III] and H β (lower panel) versus the [Fe II] equivalent width, using the data from the QSFIT catalogue.

Another important set of correlations has been highlighted by means of the principal component analysis. The spectral quantities which build up the ‘Eigenvector 1’ (i.e. the main driver for the variance in the observed spectral properties, Boroson & Green 1992) are the peak luminosity ratios between [O III] and H β, the equivalent width of the [Fe II] complex at optical wavelengths, the width of the H β emission line and a few others. These quantities present strong correlations, which are likely the manifestation of the underlying AGN physics (e.g. Wang, Brinkmann & Bergeron 1996; Boroson 2002; Sulentic et al. 2003). In Fig. 16, we show two such correlations between the width of the broad H β emission line (upper panel) and the peak luminosity ratios between [O III] and H β (lower panel) versus the [Fe II] equivalent width, using the data from the QSFIT catalogue.

These examples show that the data in the QSFIT catalogue allow us to recover the most common empirical relations known in literature. By exploiting the very large QSFIT catalogue and its additional estimates not previously available (such as the AGN continuum slopes or the host galaxy luminosities), it will be possible to study in a much deeper detail the correlations. These topics will be the subject of a forthcoming paper.

5.2 Reliability of continuum slope estimates

In order to assess the reliability of QSFIT continuum estimates, we compared the SDSS de-reddened averaged (or composite) spectrum with the average spectrum of the QSFIT continuum components in a few subsamples spanning a very narrow redshift range. Each

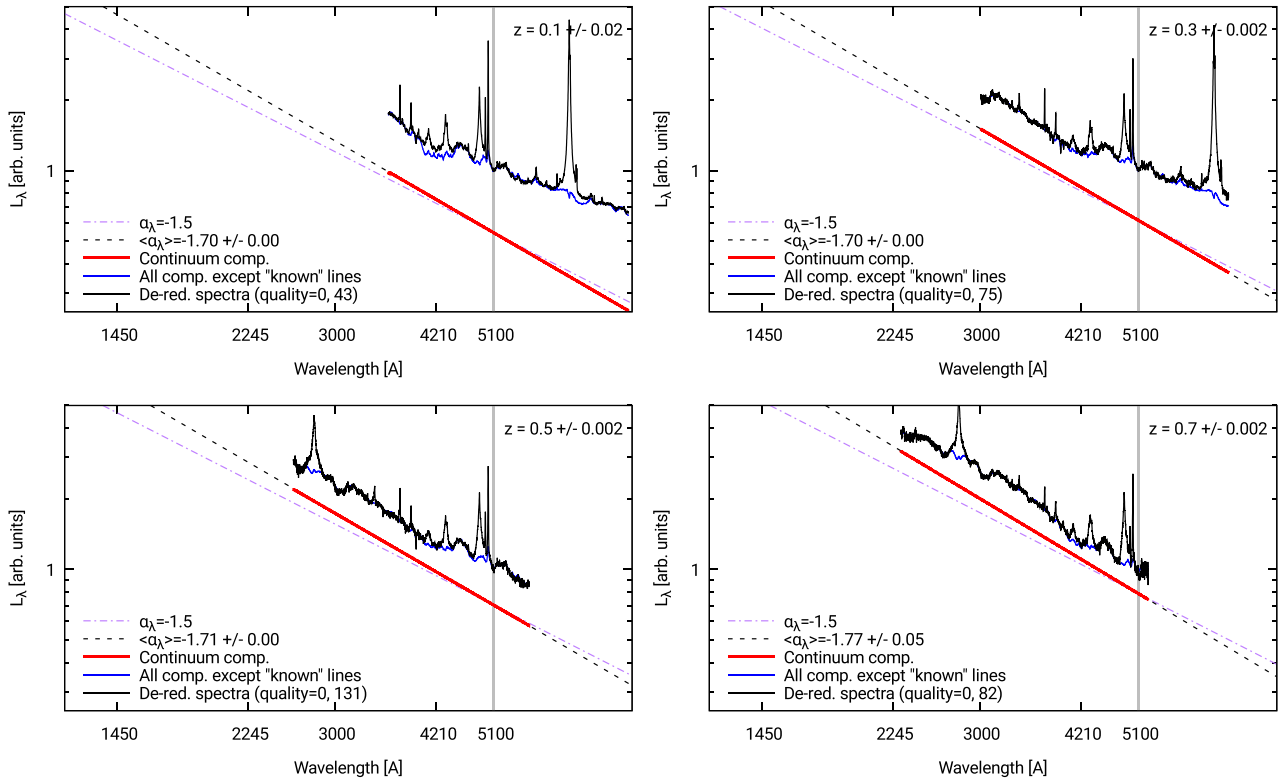


Figure 17. Comparison of SDSS de-reddened spectra and QSFIT components composites in very narrow redshift bins (shown in the upper right corner of each plot). All composite spectra are calculated as the geometrical average of the individual spectra in the subsample. Each spectrum is normalized by the source luminosity at wavelength $\lambda_{\text{norm}} = 5100 \text{ \AA}$ (vertical grey line). The composite spectrum of SDSS de-reddened spectra for all sources in the redshift bin whose continuum quality flag at λ_{norm} is 0 (CONT5100__QUALITY, Appendix A) is shown with a black solid line. The number of sources in this subsample is shown in the legend. Also shown are the composite spectrum of the QSFIT continuum component alone (red line) and the composite of the sum of all the QSFIT components except the ‘known’ emission lines (blue line). The average QSFIT slope at λ_{norm} is shown with a black dashed line, and the numerical value is shown in the legend along with the associated standard deviation of the mean. Finally, the purple dot-dashed line identifies the reference slope $\alpha_\lambda = -1.5$.

composite spectrum is calculated as the geometrical average of all spectra in the subsample, in order to preserve the slopes in a log-log plot. Before calculating the average, we normalize each de-reddened spectrum and the associated QSFIT components by the source luminosity at wavelength λ_{norm} . Fig. 17 shows such a comparison for the sources in the redshift bins $z = 0.1 \pm 0.02$, 0.3 ± 0.002 , 0.5 ± 0.002 and 0.7 ± 0.002 , respectively, using $\lambda_{\text{norm}} = 5100 \text{ \AA}$ (vertical grey line). The redshift bins are shown in the upper right corner. The bin at $z = 0.1$ is larger in order to contain a sufficient number of sources. The composite spectrum of the sources in the redshift bin whose continuum quality flag at λ_{norm} is 0 (CONT5100__QUALITY, Appendix A) is shown with a black solid line. The number of sources in this subsample is shown in the legend. Also shown are the composite spectrum of the QSFIT continuum component alone (shown with a red line) and the composite spectrum obtained by summing all the QSFIT components except the ‘known’ emission lines (blue line). The latter represents the ‘floor’ below the emission lines, i.e. the sum of the AGN, host galaxy, iron complex and Balmer continuum contributions as estimated by QSFIT. The average QSFIT slope at λ_{norm} is shown with a black dashed line, and the numerical value is shown in the legend along with the associated standard deviation of the mean. Finally, the purple dot-dashed line identifies the reference slope $\alpha_\lambda = -1.5$, to be discussed below. A similar comparison is shown in Fig. 18 for the sources at $z = 0.5 \pm 0.002$, 0.7 ± 0.002 , 1.3 ± 0.002 and 1.9 ± 0.002 , respectively, using $\lambda_{\text{norm}} = 3000 \text{ \AA}$.

In all plots, the blue line follows accurately the black line.¹⁷ Hence, QSFIT is able to find the appropriate continuum below the main emission lines, and the host galaxy, Balmer continuum and iron templates appear adequate to fit the SDSS spectra. Having considered such components in the above discussion, the remaining contribution is what we assumed to be the AGN continuum (Section 2.2), whose composite spectrum is shown with the red line, and whose average slope is shown with a black dashed line.

5.3 QSFIT catalogue main results

Here, we will discuss two of the most important results of the QSFIT analysis, namely the lack of appreciable redshift evolution of the AGN continuum slope and of the Balmer continuum luminosity (when normalized with continuum luminosity).

In Fig. 19, we show the ratio of galaxy luminosity over continuum luminosity, both estimated at $5100 \text{ \AA}$, as a function of redshift. At redshifts $\lesssim 0.3$, the AGN continuum component contribution is

¹⁷ Such agreement between the blue and black lines is best visible when the arithmetic average is used, like in Fig. 12, in place of the geometric one. However, in this section, the focus is on the reliability of the slopes hence we decided to show the geometric composites only. The plots with both geometric and arithmetic composites in all the redshift bins between 0.1 and 2 are available in the QSFIT webpage.

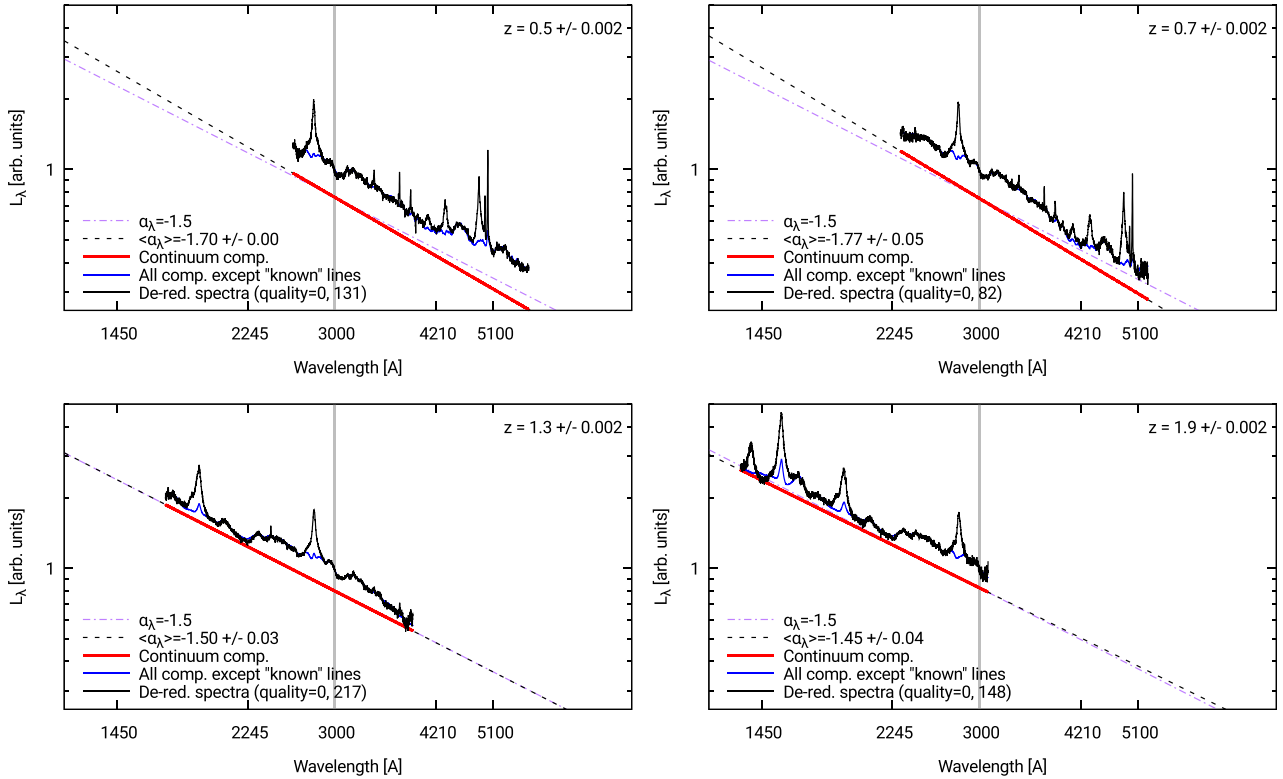


Figure 18. Comparison of SDSS de-reddened spectra and QSFIT components composites in very narrow redshift bins (shown in the upper right corner of each plot). All composite spectra are calculated as the geometrical average of the individual spectra in the subsample. Each spectrum is normalized by the source luminosity at wavelength $\lambda_{\text{norm}} = 3000 \text{ \AA}$ (vertical grey line). The composite spectrum of SDSS de-reddened spectra for all sources in the redshift bin whose continuum quality flag at λ_{norm} is 0 (CONT3000__QUALITY, Appendix A) is shown with a black solid line. The number of sources in this subsample is shown in the legend. Also shown are the composite spectrum of the QSFIT continuum component alone (red line) and the composite of the sum of all the QSFIT components except the ‘known’ emission lines (blue line). The average QSFIT slope at λ_{norm} is shown with a black dashed line, and the numerical value is shown in the legend along with the associated standard deviation of the mean. Finally, the purple dot-dashed line identifies the reference slope $\alpha_\lambda = -1.5$.

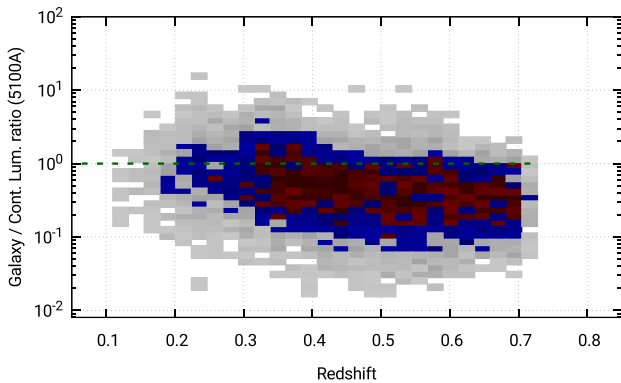


Figure 19. Ratio of galaxy luminosity over continuum luminosity, both estimated at $5100 \text{ \AA}$, as a function of redshift. At $z \lesssim 0.3$, almost half of the sources have ratio ~ 1 . At $z \gtrsim 0.5$, the galaxy contribution is typically smaller than the continuum one. The galaxy contribution is fitted up to $z = 0.8$, but this plot stops at smaller redshifts, since the $5100 \text{ \AA}$ is available only up to $z \sim 0.7$.

smaller than, or comparable with, the emission from the galaxy (Fig. 19).

This explains why our continuum luminosity estimates are smaller than the S11 ones (see Fig. 10) at low luminosities. How-

ever, for these sources, we can only indirectly probe the AGN continuum, and our estimates rely on the assumptions that the host galaxy template and the fixed continuum slope ($\alpha_\lambda = -1.7$ for the sources with $z \leq 0.6$, Section 2.2) are suitable for the source being analysed. Hence, the QSFIT continuum properties (especially the slopes) for sources with $z \lesssim 0.3$ may be biased by our assumptions. At redshifts $z \gtrsim 0.5$, the galaxy contribution become gradually less important and the AGN continuum can in principle be directly probed. However, the presence of a very broad emission feature between ~ 2200 and $\sim 4000 \text{ \AA}$ known as ‘small blue bump’ (SBB, due to the Balmer continuum and many blended iron lines, Wills et al. 1985), clearly visible in the lower left panel of Fig. 18, may introduce a bias in the QSFIT slope estimates. The bias is due to the fact that at $z \lesssim 0.8$ the SBB is not entirely visible using just the data at optical wavelengths, and consequently the AGN continuum component is ‘pulled upward’ at short wavelengths to fit the regions close to the Mg II emission line. The slope α_λ would be, therefore, systematically underestimated. As a consequence, we had to fix the continuum slope $\alpha_\lambda = -1.7$ for the sources with $z \leq 0.6$. In order to remove this limit additional data at UV wavelengths are required.

At redshifts $z \gtrsim 0.6$, the SBB is almost always entirely visible (see Fig. 18, upper right panel) and the host galaxy is typically negligible, hence the QSFIT slope estimates are not affected by the aforementioned bias, and the slope parameter is free to vary in the fit.

The 2D histograms of slope versus redshift for all sources whose `CONT*__QUALITY` is 0 (Appendix A) at the fixed rest-frame wavelengths 1450, 2245, 3000, 4210 and 5100 Å is shown in Fig. 20 (the meaning of the 2D histogram colour code is described in Section 3). The dashed red lines is at $\alpha_\lambda = -1.5$. Fig. 21 shows the same data, but considering only the sources in very narrow redshift bins ($\Delta z < 0.005$).

The slopes span the very narrow range between -1.45 and -1.65 at all redshifts between 0.8 and 2, providing remarkable evidence for the similarity of continuum slopes across cosmic time. This is in agreement with previous works: the typical value of the spectral slope at optical/UV wavelengths is $\alpha_\lambda \sim -1.5$ (e.g. Richards et al. 2006; Vanden Berk et al. 2001; Shankar et al. 2016), however our analysis allows for the first time to estimate the slopes on a very large sample, using a completely automatic procedure and just the SDSS data at optical wavelengths. In particular, our estimates are similar to those of Davis, Woo & Blaes (2007), who found $\alpha_\lambda = -1.41$ (in the wavelength range 1450–2200 Å) and $\alpha_\lambda = -1.63$ (in the wavelength range 2200–4000 Å) with rather broad distributions (see their fig. 2). At $z \gtrsim 1.1$, the slopes are remarkably constant and show no trend with redshift up to $z \sim 2$. The scatter is larger than the standard deviation of the mean in the redshift bins, hence the values are affected by small ($\lesssim 0.1$) slope systematics due to the fact that at different redshift, we observe different parts of the rest-frame spectrum. Moreover, note that the typical uncertainty associated with our estimates of α_λ (2245 Å) ranges between 0.02 and 0.1 (Fig. 22), while the widths in the distributions of Fig. 20 are significantly larger ($\sigma = 0.4$). Hence, these distributions are not significantly broadened by the random uncertainties in the fitting process. We conclude that the average continuum slope is $\alpha_\lambda \sim -1.55$, but there are AGNs whose slope is significantly smaller or larger than the average.

At $0.6 \lesssim z \lesssim 1.1$, the slopes are slightly smaller. Currently, we are not able to discriminate whether the SBB bias is the only cause of such low slope estimates, or rather if the slopes actually are smaller at $z \lesssim 0.8$ with respect to higher redshifts. Note that both the putative galaxy bias and the systematic SBB bias are not due to limitations of QSFIT: they are a consequence of having considered just the spectra at optical wavelengths rather than considering data at shorter wavelengths. By enlarging the spectral coverage, we could get rid of such biases and provide stronger constraints on the actual continuum shape at $z \lesssim 0.8$. On the other hand, using just the SDSS data allowed us to analyse a very large and homogeneous sample.

Similar conclusions applies to the ratio of the Balmer luminosity over the AGN continuum luminosity, both estimated at 3000 Å. Fig. 23 shows distribution of the Balmer continuum normalization parameter (Section 2.3) in the upper panel, and the distribution as a function of redshift in the lower panel. The distribution is quite flat at all redshifts and has a median value of 0.15 and a width of 0.3 dex.

6 CONCLUSIONS

We presented QSFIT (version 1.2), an IDL package able to automatically analyse the SDSS optical spectra of Type 1 AGNs. QSFIT is released as free software, and can be customized to analyse the spectra from other sources, or to customize the fitting recipe. QSFIT provides luminosity estimates for the AGN continuum, the Balmer continuum, both optical and UV iron blended complex, host galaxy and emission lines, as well as width, velocity offset and equivalent width of 20 emission lines, as well as several ‘quality flags’ to assess the reliability of the results.

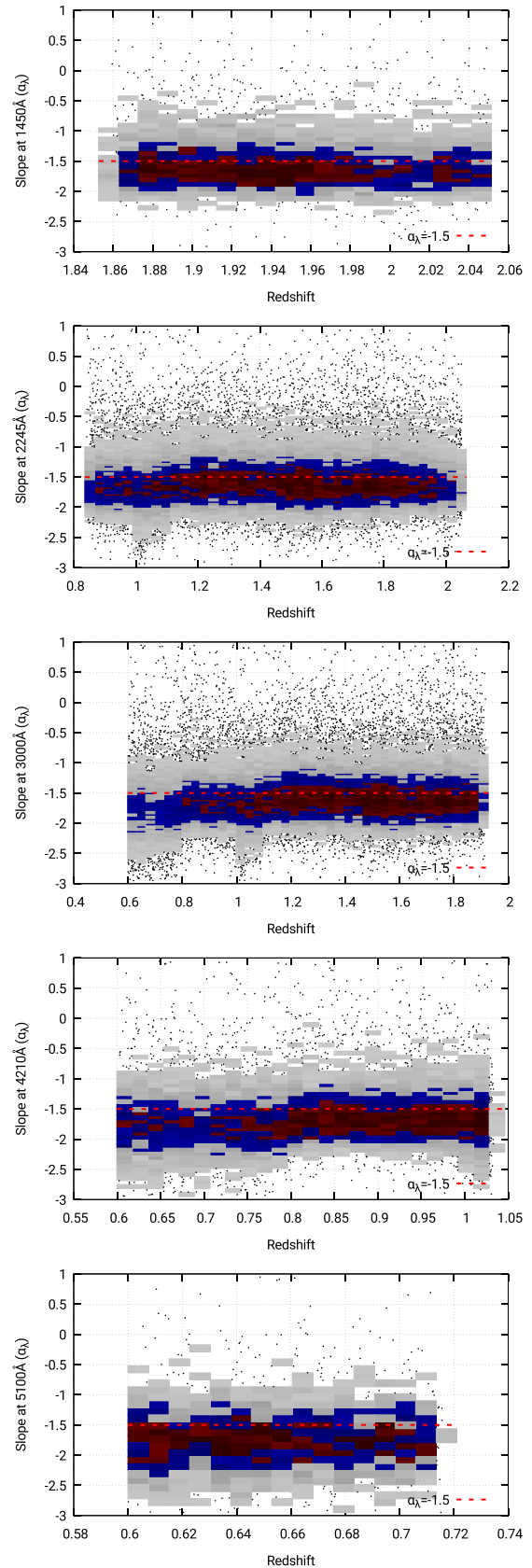


Figure 20. 2D histogram of the broad-band continuum slopes versus redshift at the fixed rest-frame wavelengths 1450, 2245, 3000, 4210 and 5100 Å (the meaning of the 2D histogram colour code is described in Section 3). The dashed red lines is at $\alpha_\lambda = -1.5$.

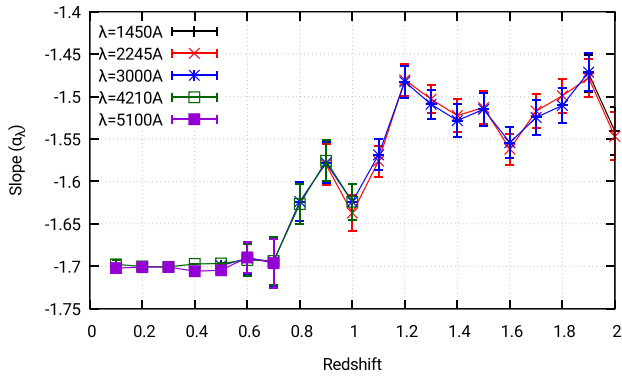


Figure 21. AGN continuum slopes as a function of redshift. The slopes at $z \leq 0.6$ at fixed $\alpha_\lambda = -1.7$

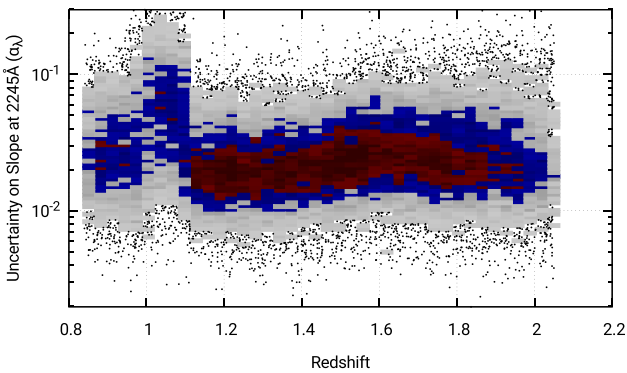


Figure 22. 2D histogram of the 1σ uncertainties in the $\alpha_\lambda(2245\text{Å})$ slope estimates versus redshift.

We used QSFIT to analyse 71 251 SDSS spectra of Type 1 AGN at $z < 2$ and to produce a publicly available catalogue of AGN spectral properties (available on the QSFIT website, along with additional data and plots). Also, we compared our results with those from the S11 catalogue (Shen et al. 2011). The quantities are in very good agreement, except for: (i) the continuum luminosity at 5100 Å for sources at low redshift, where we also considered the host galaxy contribution and (ii) the emission-line widths whose values depend on the adopted fitting algorithm for line decomposition.

We also checked the reliability of the QSFIT continuum slope estimates by a comparison with composite spectra of the observed SDSS data, and found that QSFIT performs particularly well at $z \gtrsim 0.8$, providing (for the first time) the AGN continuum slope estimates on a very large sample, using just the data at optical wavelengths and a completely automatic procedure. Also, we have shown that the continuum slope and the Balmer continuum luminosity distributions do not show any evident trend with redshift.

The release of the QSFIT code and the associated catalogue discussed in this work are the first attempt towards a new approach in analysing AGN spectral data, where all interested astronomers are given the possibility to study and compare the analysis methods, easily replicate the data analysis (even on very large samples) and possibly improve the recipes in a collaborative effort.

ACKNOWLEDGEMENTS

We gratefully acknowledge our referee, Yue Shen, for useful suggestions and comments that greatly improved the presentation of this work.

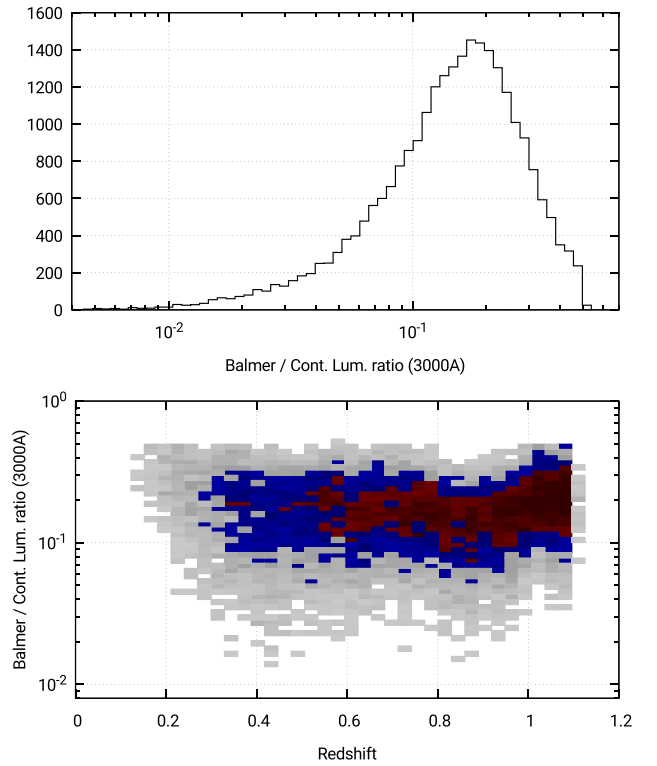


Figure 23. Upper panel: distribution of the Balmer continuum normalization parameter (Section 2.3). Lower panel: same distribution as a function of redshift. The distribution is quite flat at all redshifts and has a median value of 0.15 and a width of 0.3 dex.

We acknowledge M. Vestergaard and B.J. Wilkes for having provided the UV iron templates, and M.-P. Véron-Cetty, M. Joly and P. Véron for the optical iron templates used in the fitting process.

We warmly thank a number of undergraduate students at the University of Southampton who have widely tested QSFIT on thousands of different sources and under different input assumptions.

We also thank Chiara Moretti for her contribution to the development of the QSFIT website, and Guido Cupani for insightful discussions on fitting algorithms.

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society and the Higher Education Funding Council for England. The SDSS website is <http://www.sdss.org/>.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy, the Max-Planck-Institute for Astrophysics, New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory and the University of Washington.

REFERENCES

- Ahn C. P. et al., 2014, *ApJS*, 211, 17
- Andrae R., 2010, preprint ([arXiv:1009.2755](https://arxiv.org/abs/1009.2755))
- Baldwin J. A., 1977, *ApJ*, 214, 679
- Baldwin J. A., Wampler E. J., Gaskell C. M., 1989, *ApJ*, 338, 630
- Barth A. J. et al., 2015, *ApJS*, 217, 26
- Boroson T. A., 2002, *ApJ*, 565, 78
- Boroson T. A., Green R. F., 1992, *ApJS*, 80, 109
- Burtscher L., Jaffe W., Raban D., Meisenheimer K., Tristram K. R. W., Röttgering H., 2009, *ApJ*, 705, L53
- Calderone G., Sbarrato T., Ghisellini G., 2012, *MNRAS*, 425, L41
- Calistro Rivera G., Lusso E., Hennawi J. F., Hogg D. W., 2016, *ApJ*, 833, 98
- Cardelli J. A., Clayton G. C., Mathis J. S., 1989, *ApJ*, 345, 245
- Cristiani S., Vio R., 1990, *A&A*, 227, 385
- Davis S. W., Woo J.-H., Blaes O. M., 2007, *ApJ*, 668, 682
- Dietrich M., Appenzeller I., Vestergaard M., Wagner S. J., 2002, *ApJ*, 564, 581
- Elvis M. et al., 1994, *ApJS*, 95, 1
- Floyd D. J. E., Kukula M. J., Dunlop J. S., McLure R. J., Miller L., Percival W. J., Baum S. A., O’Dea C. P., 2004, *MNRAS*, 355, 196
- Francis P. J., Hewett P. C., Foltz C. B., Chaffee F. H., Weymann R. J., Morris S. L., 1991, *ApJ*, 373, 465
- Grandi S. A., 1982, *ApJ*, 255, 25
- Green P. J., Forster K., Kuraszkiewicz J., 2001, *ApJ*, 556, 727
- Grupe D., Komossa S., Leighly K. M., Page K. L., 2010, *ApJS*, 187, 64
- Heavens A., 2009, preprint ([arXiv:0906.0664](https://arxiv.org/abs/0906.0664))
- Kellermann K., Sramek R., Schmidt M., Shaffer D. B., Green R., 1989, *AJ*, 98, 1195
- Khachikian E. Y., Weedman D. W., 1974, *ApJ*, 192, 581
- Kinney A. L., Rivolo A. R., Koratkar A. P., 1990, *ApJ*, 357, 338
- Kriss G., 1994, in Crabtree D. R., Hanisch R. J., Barnes J., eds, *ASP Conf. Ser. Vol. 60, Astronomical Data Analysis Software and Systems III, Fitting Models to UV and Optical Spectral Data*. Astron. Soc. Pac., San Francisco, p. 437
- Markwardt C. B., 2009, in Bohlender D. A., Durand D., Dowler P., eds, *ASP Conf. Ser. Vol. 411, Astronomical Data Analysis Software and Systems XVIII, Non-linear Least-squares Fitting in IDL with MPFIT*. Astron. Soc. Pac., San Francisco, p. 251
- Matsuoka Y. et al., 2015, *ApJ*, 811, 91
- O’Donnell J. E., 1994, *ApJ*, 422, 158
- Pogge R. W., Peterson B. M., 1992, *AJ*, 103, 1084
- Polletta M. et al., 2007, *ApJ*, 663, 81
- Richards G. T. et al., 2006, *ApJS*, 166, 470
- Sanders D. B., Phinney E. S., Neugebauer G., Soifer B. T., Matthews K., 1989, *ApJ*, 347, 29
- Schneider D. P. et al., 2010, *AJ*, 139, 2360
- Shankar F. et al., 2016, *ApJ*, 818, L1
- Shen Y., Liu X., 2012, *ApJ*, 753, 125
- Shen Y. et al., 2011, *ApJS*, 194, 45
- Silva L., Granato G. L., Bressan A., Danese L., 1998, *ApJ*, 509, 103
- Storey P. J., Hummer D. G., 1995, *MNRAS*, 272, 41
- Sulentic J. W., Marziani P., Dultzin-Hacyan D., 2000, *ARA&A*, 38, 521
- Sulentic J. W., Zamfir S., Marziani P., Bachev R., Calvani M., Dultzin-Hacyan D., 2003, *ApJ*, 597, L17
- Telfer R. C., Zheng W., Kriss G. A., Davidsen A. F., 2002, *ApJ*, 565, 773
- Urry C. M., Padovani P., 1995, *PASP*, 107, 803
- Vanden Berk D. E. et al., 2001, *AJ*, 122, 549
- Véron-Cetty M.-P., Joly M., Véron P., 2004, *A&A*, 417, 515
- Vestergaard M., Wilkes B. J., 2001, *ApJS*, 134, 1
- Vestergaard M., Fan X., Tremonti C. A., Osmer P. S., Richards G. T., 2008, *ApJ*, 674, L1
- Wang T., Brinkmann W., Bergeron J., 1996, *A&A*, 309, 81
- Wills B. J., Netzer H., Wills D., 1985, *ApJ*, 288, 94
- York D. G. et al., 2000, *AJ*, 120, 1579

APPENDIX A: DESCRIPTION OF THE QSFIT CATALOGUE

The QSFIT (version 1.2) catalogue is deployed as a FITS table with 279 columns and 71 251 rows, available at the following address:

http://qsfit.inaf.it/cat_1.0/fits/qsfit_1.0.fits

On the QSFIT website you will also find a second version of the catalogue where we added, for each analysed source, the corresponding data from the S11 catalogue to allow an easy comparison. This latter FITS file has 434 columns.

In this section, we will list the name and meaning of each column in the QSFIT catalogue. The columns whose name has a `__QUALITY` suffix require a preliminary explanation: they store the ‘quality flags’, whose purpose is to characterize the reliability of the associated quantity. If a quality flag is raised for a given source, then the corresponding quantity should be used cautiously, and a further analysis is typically required. To save space in the final catalogue, these flags are coded as ‘bits’, i.e. quantities which can be either 0 (flag not raised) or 1 (flag raised), and packed into a binary representation of an integer number (the bitmask). For instance, a quality flag equal to 13 means that the bits 0, 2, 3 are raised, since $2^0 + 2^2 + 2^3 = 13$. A quality flag equal to 0 means that no flag has been raised and the associated quantity is considered reliable (within the capabilities of QSFIT). The meaning of each bit in the quality flag is discussed below.

The list of columns provided in the QSFIT catalogue is:

- (i) SPEC: name of the SDSS-DR10 FITS file containing the spectrum;
- (ii) SDSS_NAME: the SDSS name of the source (J2000.0);
- (iii) RA: the right ascension (J2000.0) of the source;
- (iv) DEC: the declination (J2000.0) of the source;
- (v) REDSHIFT: the redshift of the source;
- (vi) PLATE: spectroscopic plate number;
- (vii) FIBER: spectroscopic fibre number;
- (viii) MJD: Modified Julian Day of spectroscopic observation;
- (ix) E_BV: the colour excess used to de-redden the SDSS spectrum;
- (x) NDATA: the number of channels in the spectrum used in the analysis (Section 2.8.1);
- (xi) GOOD_FRACTION: the fraction of ‘good’ channels (whose SDSS mask is 0) over the total number of channels. This quantity is ≥ 0.75 for all sources analysed in this work (Section 2.8.1);
- (xii) MEDIAN_FLUX, MEDIAN_ERR: the median flux and median uncertainty in the input spectrum, in units of 10^{-17} erg s $^{-1}$ cm $^{-2}$ Å $^{-1}$;
- (xiii) GALAXY__LUM, GALAXY__LUM_ERR: the host galaxy νL_ν luminosity at 5500 Å and its uncertainty, in units of 10^{42} erg s $^{-1}$ (Section 2.4);
- (xiv) GALAXY__QUALITY: quality flag for the galaxy estimates. The meaning of the bits are:
 - (a) Bit 0: fit of galaxy template is not sensible at this redshift;
 - (b) Bit 1: either the luminosity or its uncertainty are NaN or equal to zero;
 - (c) Bit 2: luminosity relative uncertainty > 1.5;
- (xv) CONT*__WAVE : where ‘*’ is a number in the range 1–5: wavelengths in Å where the continuum luminosity and slopes (see below) has been computed. There are five such quantities whose values are equally spaced in the logarithmic wavelength range [$\lambda_{\min}$, $\lambda_{\max}$];

(xvi) `CONT*__LUM`, `CONT*__LUM_ERR`: where ‘*’ is in the range 1–5, or equal to 1450, 2245, 3000, 4210 or 5100: AGN continuum (Section 2.2) νL_ν luminosities and their uncertainties, in units of $10^{42} \text{ erg s}^{-1}$, estimated at wavelengths `CONT*__WAVE` or at the fixed rest-frame wavelengths 1450, 2245, 3000, 4210 or 5100 Å;

(xvii) `CONT*__SLOPE`, `CONT*__SLOPE_err`: where ‘*’ is in the range 1–5, or equal to 1450, 2245, 3000, 4210 or 5100: AGN continuum slopes and their uncertainties estimated at wavelengths `CONT*__WAVE` or at the fixed rest-frame wavelengths 1450, 2245, 3000, 4210 or 5100 Å. The slopes are calculated as the derivative of the continuum component in the $\log \lambda$ – $\log L_\lambda$ plane, this approach allows us to consider the most general case of the continuum modelled as a smoothly broken power law (Appendix B5). All the uncertainties are identical, since they are given by the uncertainties of the α_λ and $\Delta\alpha_\lambda$ parameters of the AGN continuum component (Section 2.2). Note that for sources with $z < 0.6$, the `CONT*__SLOPE` is always -1.7 and the `CONT*__SLOPE_err` is always 0;

(xviii) `CONT*__GALAXY` where ‘*’ is in the range 1–5, or equal to 1450, 2245, 3000, 4210 or 5100: host galaxy νL_ν luminosities, in units of $10^{42} \text{ erg s}^{-1}$, estimated at wavelengths `CONT*__WAVE` or at the fixed rest-frame wavelengths 1450, 2245, 3000, 4210 or 5100 Å. These quantities are provided for a quick comparison with the AGN continuum luminosities. The relative uncertainty of `CONT*__GALAXY` can be estimated as `GALAXY__LUM_ERR / GALAXY__LUM`;

(xix) `CONT*__QUALITY`: where ‘*’ is in the range 1–5, or equal to 1450, 2245, 3000, 4210 or 5100: quality flag for the AGN continuum estimates. The meaning of the bits are:

- (a) Bit 0: wavelength is outside the observed range;
- (b) Bit 1: either the luminosity or its uncertainty are NaN or equal to zero;
- (c) Bit 2: luminosity relative uncertainty > 1.5 ;
- (d) Bit 3: either the slope or its uncertainty are NaN or equal to zero;
- (e) Bit 4: slope hits a limit in the fit;
- (f) Bit 5: slope uncertainty > 0.3 ;

(xx) `IRONUV__LUM`, `IRONUV__LUM_ERR`: the integrated luminosity and its uncertainty for the iron template at UV wavelengths, in units of $10^{42} \text{ erg s}^{-1}$ (Section 2.5.2);

(xxi) `IRONUV__FWHM`, `IRONUV__FWHM_ERR`: the FWHM and its uncertainty of the Gaussian profile used to broaden the template, in units of km s^{-1} . The SDSS data cannot reliably constrain this quantity, hence `IRONUV__FWHM` is always 3000 km s^{-1} and `IRONUV__FWHM_ERR` is always 0;

(xxii) `IRONUV__EW`, `IRONUV__EW_ERR`: the equivalent width and its uncertainty for the iron template at UV wavelengths, in units of Å;

(xxiii) `IRONUV__QUALITY`: quality flag for the UV iron template. The meaning of the bits are:

- (a) Bit 0: fit of iron template is not sensible at this redshift;
- (b) Bit 1: either the luminosity or its uncertainty are NaN or equal to zero;
- (c) Bit 2: luminosity relative uncertainty > 1.5 ;

(xxiv) `IRONOPT_BR__LUM`, `IRONOPT_BR__LUM_ERR`: the integrated luminosity and its uncertainty for the ‘broad’ iron template at optical wavelengths, in units of $10^{42} \text{ erg s}^{-1}$ (Section 2.5.1);

(xxv) `IRONOPT_BR__FWHM`, `IRONOPT_BR__FWHM_ERR`: the FWHM and its uncertainty of the Gaussian profile used to

build the template, in units of km s^{-1} . The SDSS data cannot reliably constrain this quantity, hence `IRONOPT_BR__FWHM` is always 3000 km s^{-1} and `IRONOPT_BR__FWHM_ERR` is always 0;

(xxvi) `IRONOPT_BR__UNK_COUNT`: number of ‘unknown’ components (Section 2.7), whose centre lies in the range 4460–4680 Å or 5150–5520 Å, i.e. which overlaps to the optical iron template;

(xxvii) `IRONOPT_BR__UNK_LUM`, `IRONOPT_BR__UNK_LUM_ERR`: the integrated luminosity and its uncertainty for the ‘unknown’ components (Section 2.7) overlapping the optical iron template, in units of $10^{42} \text{ erg s}^{-1}$;

(xxviii) `IRONOPT_BR__EW`, `IRONOPT_BR__EW_ERR`: the equivalent width and its uncertainty for the ‘broad’ iron template at optical wavelengths, in units of Å;

(xxix) `IRONOPT_BR__QUALITY`: quality flag for the broad optical iron template. The meaning of the bits are:

- (a) Bit 0: fit of iron template is not sensible at this redshift;
- (b) Bit 1: either the luminosity or its uncertainty are NaN or equal to zero;
- (c) Bit 2: luminosity relative uncertainty > 1.5 ;

(xxx) `IRONOPT_NA__LUM`, `IRONOPT_NA__LUM_ERR`, `IRONOPT_NA__FWHM`, `IRONOPT_NA__FWHM_ERR`, `IRONOPT_NA__QUALITY`, `IRONOPT_BR__EW`, `IRONOPT_BR__EW_ERR`: same quantities as above, but for the ‘narrow’ optical template (Section 2.5.1);

(xxxi) `*__NCOMP`: where ‘*’ is one of the prefix listed in Table 1: number of components used to account for the emission line. In the current QSFIT implementation, narrow lines are always modelled with a single component;

(xxxii) `*__LUM`, `*__LUM_ERR`: where ‘*’ is one of the prefix listed in Table 1: emission-line integrated luminosities and their uncertainties, in units of $10^{42} \text{ erg s}^{-1}$;

(xxxiii) `*__FWHM`, `*__FWHM_ERR`: where ‘*’ is one of the prefix listed in Table 1: emission-line FWHM and their uncertainties, in units of km s^{-1} ;

(xxxiv) `*__VOFF`, `*__VOFF_ERR`: where ‘*’ is one of the prefix listed in Table 1: emission-line velocity offsets and their uncertainties, in units of km s^{-1} . A positive value means the line is blueshifted and a negative value means the line is redshifted;

(xxxv) `*__EW`, `*__EW_ERR`: where ‘*’ is one of the prefix listed in Table 1: the equivalent width and its uncertainty in units of Å;

(xxxvi) `*__QUALITY`: where ‘*’ is one of the prefix listed in Table 1: quality flag for the emission line. The meaning of the bits are:

- (a) Bit 0: either the luminosity or its uncertainty are NaN or equal to zero;
- (b) Bit 1: luminosity relative uncertainty > 1.5 ;
- (c) Bit 2: either the FWHM or its uncertainty are NaN or equal to zero;
- (d) Bit 3: FWHM value hits a limit in the fit;
- (e) Bit 4: FWHM relative uncertainty > 2 ;
- (f) Bit 5: either the V_{off} or its uncertainty are NaN or equal to zero;
- (g) Bit 6: V_{off} value hits a limit in the fit;
- (h) Bit 7: V_{off} uncertainty $> 500 \text{ km s}^{-1}$.

(xxxvii) `LINE_HA_BASE__NCOMP`, `LINE_HA_BASE__LUM`, `LINE_HA_BASE__LUM_ERR`, `LINE_HA_BASE__FWHM`, `LINE_HA_BASE__FWHM_ERR`, `LINE_HA_BASE__VOFF`, `LINE_HA_BASE__VOFF_ERR`,

LINE_HA_BASE_EW, LINE_HA_BASE_EW_ERR, LINE_HA_BASE_QUALITY: same quantities as above, but for the H α 'base' emission line (Section 2.6);

(xxxviii) BALMER_LUM, BALMER_LUM_ERR: vL_v luminosity and its uncertainty for the Balmer continuum component (Section 2.3) estimated at 3000 Å, in units of 10^{42} erg s $^{-1}$;

(xxxix) BALMER_RATIO, BALMER_RATIO_ERR: luminosity ratio of the high-order blended line complex to the Balmer continuum (see Fig. 1);

(xl) BALMER_LOGT, BALMER_LOGT_ERR: logarithm of electron temperature and its uncertainty for the Balmer component in units of log K. The SDSS data cannot reliably constrain this quantity, hence BALMER_LOGT is always log (15 000) K and BALMER_LOGT_ERR is always 0;

(xli) BALMER_LOGNE, BALMER_LOGNE_ERR: logarithm of electron density and its uncertainty for the Balmer component in units of log cm $^{-3}$. The SDSS data cannot reliably constrain this quantity, hence BALMER_LOGNE is always 9 and BALMER_LOGNE_ERR is always 0;

(xlii) BALMER_LOGTAU, BALMER_LOGTAU_ERR: logarithm of the optical depth, and its uncertainty, at the Balmer edge. The SDSS data cannot reliably constrain this quantity, hence BALMER_LOGTAU and BALMER_LOGTAU_ERR are always 0;

(xliii) BALMER_FWHM, BALMER_FWHM_ERR: FWHM and its uncertainty of the Gaussian profile used to broaden the Balmer component, in units of km s $^{-1}$. The SDSS data cannot reliably constrain this quantity, hence BALMER_FWHM is always 5000 km s $^{-1}$ and BALMER_FWHM_ERR is always 0;

(xliv) BALMER_QUALITY: quality flag for the Balmer component. The meaning of the bits are:

- (a) Bit 0: fit of Balmer template is not sensible at this redshift;
- (b) Bit 1: either the luminosity or its uncertainty are NaN;
- (c) Bit 2: luminosity relative uncertainty > 1.5.

(xlv) CHISQ: the value of the χ^2_{red} , calculated using the data uncertainties provided in the SDSS FITS files;

(xlv) DOF: the number of degrees of freedom (number of data – number of free model parameters) in the fit;

(xlvii) ELAPSED_TIME: the elapsed time in seconds required to analyse the source.

All uncertainties in the QSFIT estimates are quoted at the 68 per cent level, i.e. 1σ ;

APPENDIX B: QSFIT INSTALLATION, USAGE AND CUSTOMIZATION

B1 QSFIT installation, compilation and test

The only pre-requisite to run QSFIT is IDL (ver. ≥ 8.1). If you need to view/generate the plots you should also install GNUPLOT (ver. ≥ 5.0). The IDL code of QSFIT (version 1.2) can be downloaded at the following address:

<https://github.com/gcalderone/qsfit/releases>

Once downloaded the QSFIT package should be unpacked and placed in a directory. Then, you should start an IDL session, change to the directory where you stored the unpacked files, e.g. (on Windows):

```
IDL> CD, 'C:\path\where\qsfit\is\located'
or (on Linux)
IDL> CD, '/path/where/qsfit/is/located'
```

and compile all the QSFIT procedures and dependencies by calling the compile procedure:

```
IDL> compile
```

Note that all the dependencies (the Astrolib and MPFIT routines) are included in the IDL/Contrib directory of the QSFIT package and are automatically compiled in the step above. Also note that the only required path in the IDL_PATH variable is the IDL default path to the lib directory, you do not have to add any other path to use QSFIT. Finally, you should avoid calling the compile procedure and another QSFIT routine from the same program. The best practice is to call compile manually at the very beginning of the IDL session, before any other call.

The QSFIT package already comes with an SDSS-DR 10 FITS file, suitable to immediately test the code using the following command:

```
IDL> res = qsfit('data/spec-0752-52251-0323.fits', $z=0.3806, ebv=0.06846)
```

Once the analysis is finished (it will require a few seconds), you can plot the results with the following command (GNUPLOT is required to run this step):

```
IDL> qsfit_plot, res
```

If you went through all the steps in this section without errors, then QSFIT is installed properly, and you can start the data analysis as discussed in the next section.

B2 QSFIT usage

In order to use QSFIT to perform spectral analysis, you should first download all the necessary spectra. In its current implementation (version 1.2), QSFIT supports only SDSS-DR10 FITS files of sources with $z < 2$. You can find these files using the 'Object explorer' facility on the SDSS-DR10 website, at the address:

<http://skyserver.sdss.org/dr10/en/tools/explore/obj.aspx>

Alternatively you may download a FITS file using a URL similar to the following:

<http://dr10.sdss3.org/sas/dr10/sdss/spectro/redux/26/spectra/0752/spec-0752-52251-0323.fits>

The numbers to be changed are:

- (i) the spectroscopic plate number (0752 in the example above. Note that: this number appears twice in the URL);
- (ii) the MJD of observation (52251 in the example above);
- (iii) the spectroscopic fibre number (0323 in the example above).

The numbers for a specific object may be found in the SDSS-DR10 website or in the QSFIT catalogue (columns PLATE, MJD and FIBER). The QSFIT package already comes with one such FITS file in the 'data' directory. This can be used to quickly test the QSFIT functionalities (see Appendix B1).

To run the QSFIT analysis you should simply call the qsfit function, with a command similar to the following:

```
IDL> res = qsfit('data/spec-0752-52251-0323.fits', $z=0.3806, ebv=0.06846)
```

The parameters are:

- (i) the path to a SDSS-DR10 FITS file;
- (ii) the redshift of the source;
- (iii) the $E(B - V)$ colour excess (in magnitudes).

With this command you may easily replicate all the data analysis discussed in this paper. The FITS file names, the redshifts and the colour excess used in the analysis are all available in the QSFIT catalogue (in the columns: SPEC, REDSHIFT and E_BV).

The QSFIT analysis will require a few seconds (see Fig. 9), and all the results will be returned in the `res` variable, as a nested structure. Such structure is not suitable to be collected in a catalogue like the QSFIT one (Section 3), hence we provide a procedure to re-arrange the nested structure into a ‘flattened’ one, i.e. a structure whose tags are all scalar values. To obtain a flattened structure, you should use a command like the following:

```
IDL> flat = qsfit_flatten_results(res)
```

The tags in the `flat` structure are described in Appendix A, and the QSFIT catalogue (Section 3) is actually a collection of QSFIT flattened structures. The nested structure is described in further detail in Appendix B3.

If you want the results to be permanently saved in a directory (which must have been previously created), use the `OUTDIR` keyword, e.g.:

```
IDL> res = qsfit('data/spec-0752-52251-0323.fits', $z=0.3806, ebv=0.06846, $
    outdir='output')
```

This will save all the logs in a file named `output/spec-0752-52251-0323_QSFIT.log` and the results in `output/spec-0752-52251-0323_QSFIT.dat`. The latter contains a binary representation of the results, and can be used to restore the data in IDL (using the `restore` procedure). Note that if you run again the above command the program will stop immediately with the message `File output/spec-0752-52251-0323_QSFIT.dat already exists`. This is not an error, it simply says that the analysis has already been performed and there is no need to run it again. To actually re-run the analysis, you should either manually delete that file or avoid using the `OUTDIR` keyword.

Once the analysis is finished, you can plot the results with the following command (GNUPLOT is required to run this step):

```
IDL> qsfit_plot, res
```

This will open two GNUPLOT windows with the plots shown in Fig. 6. To show the data re-binned by a factor of, say, 5, you can use the following command:

```
IDL> res.gfit.plot.(0).main.rebin = 5
IDL> qsfit_plot, res
```

Note that in the commands above you should use the ‘nested’ structure (`res`), not the ‘flattened’ one (`flat`), since the latter does not contain all the necessary data to draw the plots.

You may save all the plot data in a file suitable to be later loaded in GNUPLOT, and re-directed into one of the many terminals supported by GNUPLOT. This step is accomplished by the following command:

```
IDL> qsfit_plot, res, $filename='plot/spec-0752-52251-0323'
```

This command will create two files in the `plot` directory (which must have been previously created): one with the `.gp` extension and one with the `_resid.gp` suffix. The former is the actual plot, to compare the data and the model (like the one in Fig. 6, upper panel). The latter is the plot of the residuals (Fig. 6, lower panel). These are just text files containing the data and the GNUPLOT command to actually show the plots and are completely decoupled from IDL. You can load them in GNUPLOT using the following commands:

```
gnuplot> load 'plot/spec-0752-52251-0323.gp'
gnuplot> load 'plot/spec-0752-52251-0323_resid.gp'
```

You may even modify them with a text editor before loading into GNUPLOT, in order to change the way the plots are displayed (e.g. colours, format, legends, etc.) or the terminal being used. The default GNUPLOT terminal is `wxt`, which opens an interactive window showing the plot. To obtain a PDF copy you may simply use the ‘Ex-

port to plot file’ button. Alternatively you may modify the `.gp` files by changing the terminal. E.g. replace the line `set term wxt` with:

```
set term pdf
set output 'filename.pdf'
and reload the file in GNUPLOT.
```

B3 The nested structure returned by QSFIT

The nested structure returned by the `qsfit` function contains much more information than the flattened one used to build the QSFIT catalogue. These further informations are stored in the following tags:

(i) `QSFIT_VERSION`: a string containing the version of QSFIT which generated the results;

(ii) `LOG`: an array of strings containing all the log messages generated during the analysis. To print them you may use the command `IDL> gprint, res.log`;

(iii) `CONT`: an array of structures containing all the results related to the continuum component. These informations are also contained in the flattened structure, although in a different format. The `CONT` tag is very useful to quickly print all the continuum quantities with the command `IDL> gps, res.cont`;

(iv) `LINES`: an array of structures containing all the results related to the emission-line components. These informations are also contained in the flattened structure, although in a different format. The `LINES` tag is very useful to quickly print all the emission-line quantities with the command: `IDL> gps, res.lines`;

(v) `GFIT`: this structure contains the whole state of the GFIT framework (see Appendix B6) used to actually perform the fitting: i.e. the input data, the component settings, the best-fitting values and uncertainties, the plotting settings, etc. For instance, if you want to manually plot the data and the best-fitting model (using the IDL plotting facilities) you may use the following commands:

```
IDL> plot, $res.gfit.comp.(0).x, res.gfit.
    comp.(0).y, psym=3
IDL> oplot, $res.gfit.comp.(0).x, res.gfit.
    comp.(0).m, col=255
```

Or, if you want to print the best-fitting value and uncertainty of the α_λ parameter in the continuum component (see equation B2):

```
IDL> print, a.gfit.comp.continuum.alpha1.val,
    $ a.gfit.comp.continuum.alpha1.err
```

A complete description of the GFIT structure is beyond the purpose of this paper. Interested readers can explore the structure using the `help` command, or read the documentation in the `IDL/Glib/gfit` directory.

B4 Parameter uncertainty estimates through Monte Carlo resampling

As discussed in Section 2.8.3, the QSFIT uncertainties provided by the Fisher matrix method can be considered as rough estimates, mainly because of correlation among model parameters (e.g. between continuum and galaxy luminosities, or between emission-line luminosities and widths) and asymmetric uncertainty intervals. A better and more reliable estimate for the uncertainty of each parameter is provided by the Monte Carlo resampling method (Burtscher et al. 2009; Andrae 2010), which only relies on the assumptions that the QSFIT model provide a reliable representation of the data and that the uncertainties on the observed spectrum are Gaussian distributed with known standard deviation. To apply the resampling method, we consider the best-fitting model

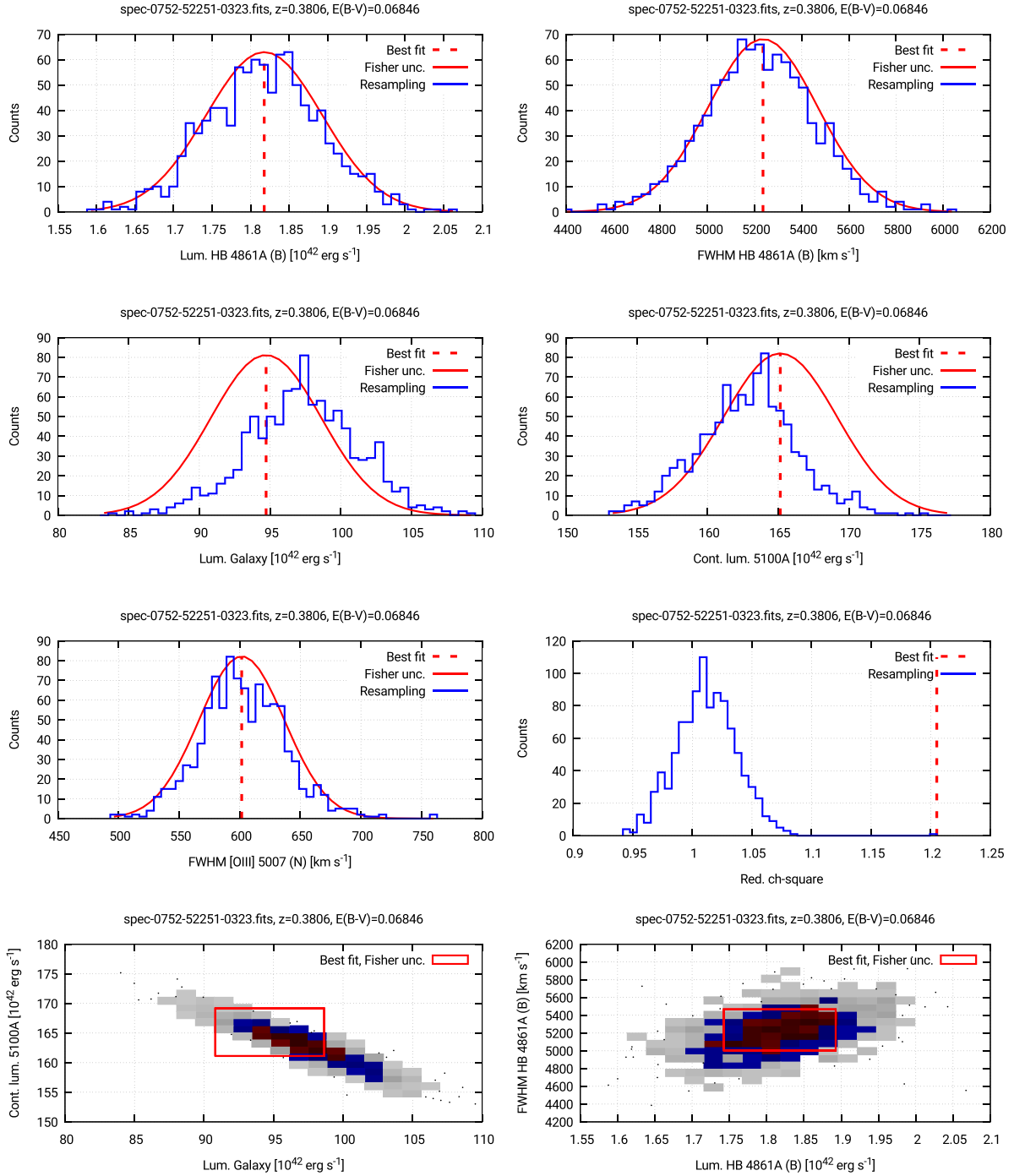


Figure B1. Comparison of expected distributions of QSFIT best-fitting parameters and associated Fisher uncertainties (red lines) and best-fitting parameters from resampled data (blue line). The sixth panel shows the distribution of the reduced χ^2 values (blue) of resampled data and the QSFIT value (red). The last two panels show the 2D distribution of two correlated parameters following the same rules described in Section 2.9. The red square identifies the region of $\pm 1\sigma$ Fisher uncertainties for both parameters.

parameters provided by QSFIT (Section 2.8.3), but neglect the associated uncertainties. Instead, we assume the best-fitting model to be the ‘true’ one, with no uncertainties, and generate a set of mock spectra whose flux, in each spectral channel, is a Gaussian random variable with standard deviation equal to the uncertainty of input data. Then we re-run the QSFIT analysis on the mock spectra and collect all the results (again, neglecting the Fisher uncertainties). Finally,

we consider the distributions of best-fitting values for each model parameter, whose width provide an estimate for the associated uncertainties. For example to estimate the 1σ uncertainty we may consider the interval between the 15.8 and 84.2 percentile of the whole distribution.

Examples of such distributions are shown in Figs B1 and B2 for the same source analysed in Fig. 6 and for spec-0433-51873-0181

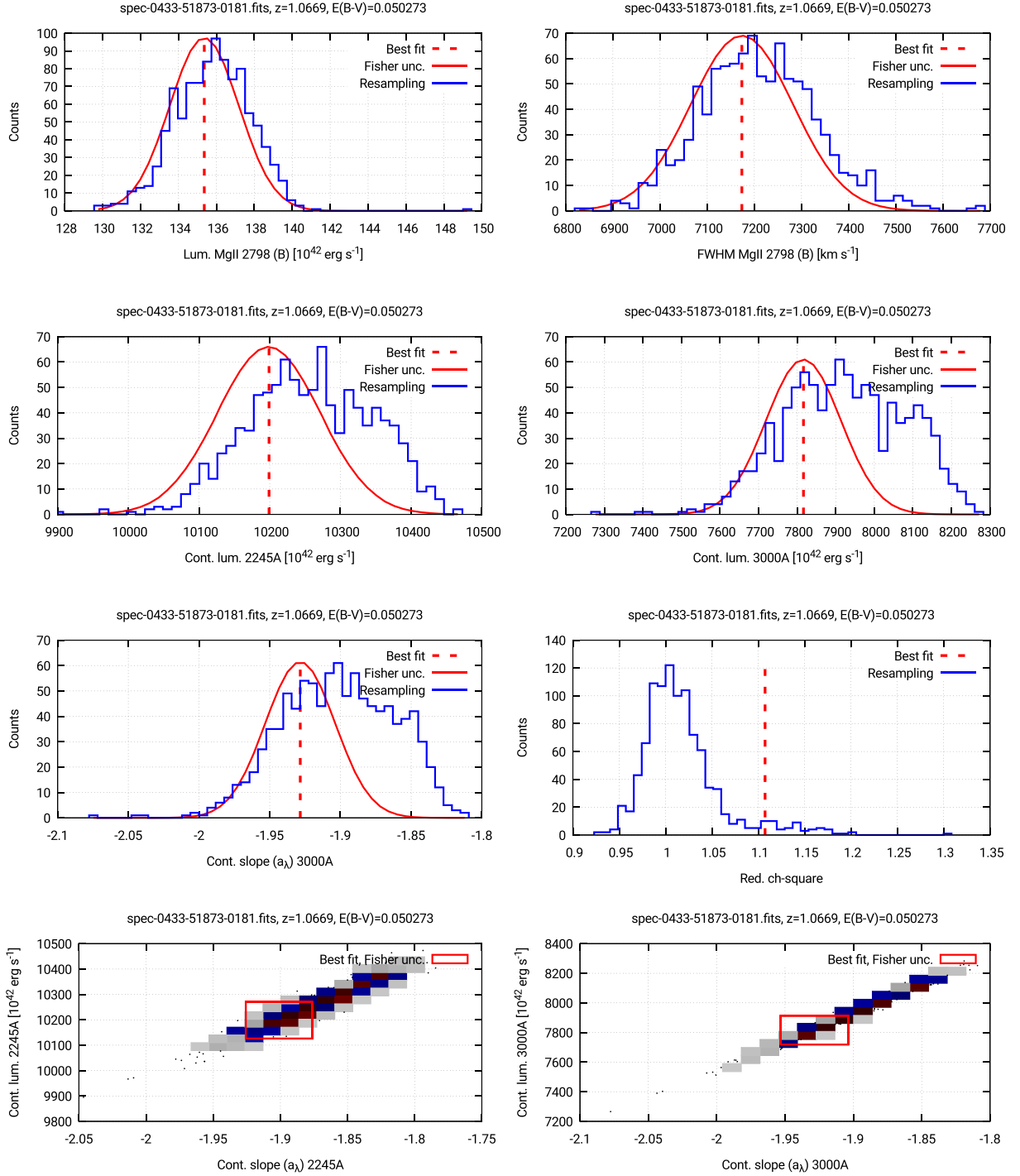


Figure B2. Same as Fig. B1, but considering the analysis of another source.

($z=1.07$), respectively. The first five histograms show the comparison between the Gaussian distribution expected from the best-fitting parameter value and its associated Fisher uncertainty (shown in red) as returned by QSFIT, and the actual distribution of best-fitting parameters when the data are resampled 10^3 times (shown in blue). The sixth panel shows the distribution of the reduced χ^2 values (in blue) which is obviously centred on one, while the QSFIT value (in red) may be slightly larger than one. The last two panels shows the

2D distribution of two correlated parameters following the same rules described in Section 3. The red square identifies the region of $\pm 1\sigma$ Fisher uncertainties for both parameters. In both Figs B1 and B2, the Fisher values provide reasonable estimates of parameter uncertainties, in a significantly smaller fraction of time with respect to Monte Carlo estimates. The latter, however, provide uncertainties which are significantly more reliable and should be used whenever the parameters of interest are correlated.

To run a Monte Carlo resampling method, a thousand times simply add the `resample=1000` keyword to the `qsfit` call (Appendix B2). The return value will be an array of structures (1000 in this case), instead of a single one.

B5 The QSFIT continuum component

The AGN continuum component used in the analysis of the catalogue (Section 3) is a simple power law. However, the QSFIT continuum component can be customized to behave as a smoothly broken power-law model, to suit those cases where the spectral coverage is sufficiently large to constrain the extra continuum parameters.

The smoothly broken power law is modelled as follows:

$$L_\lambda = A \left(\frac{\lambda}{\lambda_b} \right)^{\alpha_\lambda} \left[\frac{1 + (\lambda/\lambda_b)^{|\Delta\alpha|\gamma}}{2} \right]^{S/\gamma} \quad (\text{B1})$$

where λ_b is the break wavelength, A is the luminosity density at $\lambda = \lambda_b$, α_λ is the spectral slope at $\lambda \ll \lambda_b$, $\Delta\alpha_\lambda$ is the slope change at $\lambda \gg \lambda_b$, S is either +1 or -1, depending on the sign of $\Delta\alpha_\lambda$ and γ is the curvature parameter. When $\lambda \ll \lambda_b$ or $\lambda \gg \lambda_b$ [equation (B1)] reduces to:

$$L_\lambda = A \times \begin{cases} \left(\frac{\lambda}{\lambda_b} \right)^{\alpha_\lambda} & \lambda \ll \lambda_b \\ \left(\frac{\lambda}{\lambda_b} \right)^{\alpha_\lambda + \Delta\alpha_\lambda} & \lambda \gg \lambda_b \end{cases} \quad (\text{B2})$$

The change of slopes occurs between the wavelengths $\lambda_1 (< \lambda_b)$ and $\lambda_2 (> \lambda_b)$ such that:

$$\log_{10} \frac{\lambda_2}{\lambda_1} \sim \frac{2}{|\Delta\alpha_\lambda| \times \gamma} \quad (\text{B3})$$

hence γ controls the ‘sharpness’ of the change in slope. In the current QSFIT implementation, the A parameter is constrained to be positive; λ_b is constrained in the range $[\lambda_{\min} + 0.2 \times (\lambda_{\max} - \lambda_{\min}), \lambda_{\min} + 0.8 \times (\lambda_{\max} - \lambda_{\min})]$; α_λ is constrained in the range $[-3, 1]$ and $\Delta\alpha_\lambda$ is constrained in the range $[-0.2, 0.2]$, although the appropriate limits depends on the available spectral coverage.

B6 QSFIT design, internals and the GFIT framework

In this section, we will provide a brief description of the principles followed while developing QSFIT, its most relevant routines, and the GFIT framework (used for model evaluation). Also, we will provide the very basic information required to understand the code in the `qsfit.pro` file. This is not intended to be a complete documentation of the code, but rather a quick and very short introduction to QSFIT internals.

Unlike the field of X-ray spectral analysis, there is not yet a well established and widely accepted procedure to automatically analyse the Type 1 AGN optical spectra. This is due to the many possible ways the problem can be approached, and to the different purposes pursued by the researchers in the field. QSFIT is just a first attempt to fill this gap but, given the above premises, it is likely that the way it operates will not be suitable to fit particular needs, and likely the code will require some customization. This implies that the best way to design QSFIT is by providing a neat and short source code with extensive documentation, easy to read and customize. We went a little further and carefully separated the procedures which are rather well established and will unlikely be modified by the user, from those that are actually related to the specific task of AGN spectral fitting. The former procedures were collected in a library in the `IDL/Glib` and `IDL/qsfit/components` directories of the QSFIT package, while the latter are all contained

in a single file named `qsfit.pro` (located in the `IDL/qsfit` directory). In the vast majority of cases, this is the only file one needs to modify to customize the fitting process. As a consequence, it is actually pointless to develop a well-defined and user-friendly interface to customize the fitting process (as is the case for, e.g. XSPEC, SHERPA, etc.), since the underlying code can be modified in an unpredictable way. The drawback of this approach is that the users should become acquainted with the QSFIT source code. On the other hand, it provides the greatest flexibility to the user, since it allows us to modify every detail of the analysis and implement very complex recipes. However, in order to facilitate the most common customization tasks even for unexperienced users, we created an IDL global variable named `!QSFIT_OPT` containing the settings a user may wish to change. The default values for such settings, and their meaning, can be found in the `qsfit_prepare_options` routine in the `qsfit.pro` file. Moreover, we marked the places in the source file `qsfit.pro`, where an easy customization can be made, the with the comment `CUSTOMIZABLE`. These lines can be easily modified even without a complete understanding of the remaining code.

The QSFIT code relies on the GFIT framework to handle the model evaluation, parameter definitions and constraints, and the generations of plots to compare the data and the model. The QSFIT framework relies on the MPFIT minimization routine (Markwardt 2009), and its use allowed us to write a very neat code and to always obtain the best performances in model evaluation, regardless of the complexity of the model. The documentation for the GFIT framework can be found in the `IDL/Glib/gfit` directory. Here, we will shortly describe its usage:

(i) the GFIT software maintains its whole state in a single structure named `gfit`. The information related to the input data, the components to build the model, the parameters in each component, the parameters constraints, the plotting settings, etc., are all stored in this variable. The `gfit` structure is available in an IDL common block with the same name, hence each program unit willing to use it should state the following line at its beginning:

```
COMMON GFIT
```

(ii) the `gfit` structure is modified by the calls to one of the `gfit_*` routines, e.g. `gfit_add_data` to add a data set, `gfit_add_comp` to add a component to the model, etc.

(iii) the values in the `gfit` structure can be directly modified by the user, e.g. the mathematical expression for the model will be set into `gfit.expr.(0).model`, the limits for the host galaxy normalization parameter will be set into `gfit.comp.galaxy.norm.limits`, etc. An absorption line can easily be included in the model by setting the normalization limits to negative values, e.g.: `gfit.comp.LINE_NAME.norm.limits = [-1, 0];`

(iv) each time the model is modified (either by adding a component, by freezing or thawing a parameter), the `gfit_compile` procedure should be called. This procedure automatically generates and compile the IDL code needed to evaluate the model, ensuring the best possible performance even with very complex model definitions;

(v) the `gfit_report` and the `gfit_plot` procedures are, respectively, used to generate a report of the GFIT state, and to generate the comparison plot between the model and the data;

(vi) the `gfit_run` procedure is used to perform the model minimization through a call to `mpfit`. Once finished, the parameter values and uncertainties are updated. The latter are available in the `gfit.comp` substructure,

e.g. `gfit.comp.continuum.norm.val`, `gfit.comp.continuum.norm.err`, etc.

As discussed above, all the relevant code for QSFIT customization is located in the `qsfit.pro` file. In the following, we will briefly mention the most relevant routines:

(i) `qsfit_log`: this procedure print log messages. All messages are also stored in an internal buffer to be dumped at the end of the analysis;

(ii) `qsfit_compile`: this procedure (re-)compiles the GFIT model. It should be called each time the model is modified;

(iii) `qsfit_lineset`: this function returns the list of ‘known’ emission lines to be considered in the analysis. It can be customized to add/remove emission-line components to the model. Disabling a few lines, e.g. the less luminous ones, is the simplest way to significantly speed up the QSFIT analysis;

(iv) `qsfit_prepare`: this procedure reads an SDSS-DR 10 FITS file and loads data into GFIT internal structure (see Section 2.8.1). If you wish to read data from another source you should modify this procedure;

(v) `qsfit_run`: this procedure implements the actual QSFIT recipe for model fitting (Section 2.8.2). If you wish to modify the sequence by which each component is added you should modify this procedure;

(vi) `qsfit_reduce`: this function reduces the data in the GFIT structure (Section 2.8.3) and returns the final nested structure (Appendix B3);

(vii) `qsfit`: this is the main function to perform the QSFIT analysis, and actually the only one supposed to be directly called by the final user;

(viii) `qsfit_report`: this procedure prints the final report of the QSFIT analysis;

(ix) `qsfit_plot`: this procedure generates the plots of the results of the analysis (Appendix B2);

(x) `qsfit_version`: this function simply returns a string with the current QSFIT version, also stored in the nested structure returned by the `qsfit` function (see Appendix B3). If you modify the `qsfit.pro` file, we suggest to modify this function as well, in order to quickly identify which version generated a given result.

APPENDIX C: QUALITY FLAGS IN THE QSFIT CATALOGUE

The columns in the QSFIT catalogue whose name has a `__QUALITY` suffix report the ‘quality flags’ for the associated quantity (see Appendix A). In this section, we report both the total number and the fraction of sources (over the total number of sources in the catalogue: 71 251) which raised a flag. Also shown are the total number and the fraction of sources with no flag raised (i.e. whose `__QUALITY` entry is equal to zero), also called ‘Good’ sources with respect to a specific quantity.

The statistics for the continuum quantities are summarized in Table C1, where each of the five wavelengths (equally spaced in the logarithmic wavelength range) are shown in the upper part of the table, in the columns labelled W1...W5. The lower part of the table reports the statistics for the 1450, 2245, 3000, 4210 and 5100 Å fixed wavelengths. The statistics for the host galaxy quantities are summarized in Table C2. The statistics for the four UV iron components and for the two optical iron quantities are shown in Table C3 (upper and lower tables, respectively). Finally, the statistics for the emission-line quantities are shown in Tables C4 and C5.

Table C1. Total number and fraction of sources (over the total number of sources in the catalogue: 71 251) which raised a flag in the `CONT*__QUALITY` bitmask. The upper table shows the statistics for the five equally spaced wavelengths in the available range (see Section 2.8.3), while the lower table shows the results for the 1450, 2245, 3000, 4210 and 5100 Å fixed wavelengths. The first row in each table shows the total number and fraction of sources which raised no flag in the bitmask (‘Good’ sources with respect to the continuum quantities).

Quality	W1		W2		W3		W4		W5	
Good ^a	70 590	99.07 per cent	70 590	99.07 per cent	70 588	99.07 per cent	70 587	99.07 per cent	70 586	99.07 per cent
Bit 0 ^b	0	0.00 per cent	0	0.00 per cent	0	0.00 per cent	0	0.00 per cent	0	0.00 per cent
Bit 1 ^c	32	0.04 per cent	32	0.04 per cent	32	0.04 per cent	32	0.04 per cent	32	0.04 per cent
Bit 2 ^d	4	0.01 per cent	4	0.01 per cent	4	0.01 per cent	4	0.01 per cent	5	0.01 per cent
Bit 3 ^e	277	0.39 per cent	277	0.39 per cent	277	0.39 per cent	277	0.39 per cent	277	0.39 per cent
Bit 4 ^f	38	0.05 per cent	38	0.05 per cent	40	0.06 per cent	41	0.06 per cent	41	0.06 per cent
Bit 5 ^g	320	0.45 per cent	320	0.45 per cent	320	0.45 per cent	320	0.45 per cent	320	0.45 per cent
Bit 6 ^h	0	0.00 per cent	0	0.00 per cent	0	0.00 per cent	0	0.00 per cent	0	0.00 per cent
Quality	1450 Å		2245 Å		3000 Å		4210 Å		5100 Å	
Good ^a	6587	9.24 per cent	52 105	73.13 per cent	60 791	85.32 per cent	25 325	35.54 per cent	14 117	19.81 per cent
Bit 0 ^b	64 647	90.73 per cent	18 962	26.61 per cent	9839	13.81 per cent	45 403	63.72 per cent	56 806	79.73 per cent
Bit 1 ^c	0	0.00 per cent	0	0.00 per cent	6	0.01 per cent	32	0.04 per cent	31	0.04 per cent
Bit 2 ^d	0	0.00 per cent	0	0.00 per cent	0	0.00 per cent	4	0.01 per cent	4	0.01 per cent
Bit 3 ^e	11	0.02 per cent	123	0.17 per cent	272	0.38 per cent	179	0.25 per cent	102	0.14 per cent
Bit 4 ^f	4	0.01 per cent	21	0.03 per cent	34	0.05 per cent	18	0.03 per cent	13	0.02 per cent
Bit 5 ^g	2	0.00 per cent	41	0.06 per cent	318	0.45 per cent	297	0.42 per cent	186	0.26 per cent
Bit 6 ^h	0	0.00 per cent	0	0.00 per cent	0	0.00 per cent	0	0.00 per cent	0	0.00 per cent

Notes. ^aAll bits set to 0.

^bBit 0: wavelength is outside the observed range.

^cBit 1: either the luminosity or its uncertainty are NaN or equal to zero.

^dBit 2: luminosity relative uncertainty > 1.5.

^eBit 3: either the slope or its uncertainty are NaN or equal to zero.

^fBit 4: slope hits a limit in the fit.

^gBit 5: slope uncertainty > 0.3.

Table C2. Total number and fraction of sources (over the total number of sources in the catalogue: 71 251) which raised a flag in the `GALAXY__QUALITY` bitmask. The first row shows the total number and fraction of sources which raised no flag in the bitmask ('Good' sources with respect to the host galaxy quantities).

Quality	Number of sources	
Good ^a	13 015	18.3 per cent
Bit 0 ^b	53 887	75.6 per cent
Bit 1 ^c	3727	5.2 per cent
Bit 2 ^d	622	0.9 per cent

Notes. ^aAll bits set to 0.

^bBit 0: fit of galaxy template is not sensible at this redshift.

^cBit 1: either the luminosity or its uncertainty are NaN or equal to zero.

^dBit 2: luminosity relative uncertainty > 1.5.

Table C3. Total number and fraction of sources (over the total number of sources in the catalogue: 71 251) which raised a flag in the `IRONUV__QUALITY` bitmasks (upper table), or in any of the `IRONOPT_BR__QUALITY` and `IRONOPT_NA__QUALITY` bitmasks (lower table). The first row in each table shows the total number and fraction of sources that raised no flag in the bitmask ('Good' sources with respect to the iron quantities).

Quality	Iron UV			
Good ^a	46 540	65.3 per cent		
Bit 0 ^b	0	0.0 per cent		
Bit 1 ^c	24 676	34.6 per cent		
Bit 2 ^d	35	0.0 per cent		
Quality	Broad		Narrow	
Good ^a	21 449	30.1 per cent	12 070	16.9 per cent
Bit 0 ^b	48 416	68.0 per cent	48 416	68.0 per cent
Bit 1 ^c	871	1.2 per cent	7466	10.5 per cent
Bit 2 ^d	515	0.7 per cent	3299	4.6 per cent

Notes. ^aAll bits set to 0.

^bBit 0: fit of iron template is not sensible at this redshift.

^cBit 1: either the luminosity or its uncertainty are NaN or equal to zero.

^dBit 2: luminosity relative uncertainty > 1.5.

Table C4. Total number and fraction of sources (over the total number of sources reported in the row labelled ‘Total’) whose emission lines were modelled with either one component ($N_{\text{comp}} = 1$), two components ($N_{\text{comp}} = 2$), etc. The line labelled ‘Total’ reports the total number of sources for which the $|\text{*_NCOMP}|$ is not equal to zero, i.e. for the sources which are in the appropriate redshift range for the emission line to be detectable (see Section 2.6). In the second part of the tables, we show the total number and fraction of sources which raised a flag in any of the emission line *_QUALITY bitmasks. The fractions refer always to the row labelled ‘Total’. Finally, the lines labelled ‘Good’ shows the total number and fraction of sources which raised no flag in the bitmask (‘Good’ sources with respect to the emission-line quantities).

Line	Si IV (B)		C IV (B)		C III (B)		Mg II (B)		Ne VI (N)	
Wavelength (Å)	1400		1549		1909		2798		3426	
$N_{\text{comp}}=1$	7375	(97.1 per cent)	12 001	(58.7 per cent)	40 138	(93.1 per cent)	48 826	(74.1 per cent)	49 970	(100.0 per cent)
$N_{\text{comp}}=2$	220	(2.9 per cent)	7709	(37.7 per cent)	2916	(6.8 per cent)	16 397	(24.9 per cent)		
$N_{\text{comp}}=3$	1	(0.0 per cent)	685	(3.4 per cent)	49	(0.1 per cent)	661	(1.0 per cent)		
$N_{\text{comp}}=4$			39	(0.2 per cent)	1	(0.0 per cent)	14	(0.0 per cent)		
$N_{\text{comp}}=5$			2	(0.0 per cent)						
$N_{\text{comp}}=6$			1	(0.0 per cent)						
$N_{\text{comp}}=7$			2	(0.0 per cent)						
Total	7596		20 439		43 104		65 898		49 970	
Quality										
Good ^a	5151	(67.8 per cent)	15 193	(74.3 per cent)	25 208	(58.5 per cent)	52 399	(79.5 per cent)	11 134	(22.3 per cent)
Bit 0 ^b	99	(1.3 per cent)	67	(0.3 per cent)	94	(0.2 per cent)	49	(0.1 per cent)	4073	(8.2 per cent)
Bit 1 ^c	62	(0.8 per cent)	53	(0.3 per cent)	87	(0.2 per cent)	86	(0.1 per cent)	4373	(8.8 per cent)
Bit 2 ^d	565	(7.4 per cent)	930	(4.6 per cent)	6493	(15.1 per cent)	654	(1.0 per cent)	22 811	(45.6 per cent)
Bit 3 ^e	661	(8.7 per cent)	861	(4.2 per cent)	4357	(10.1 per cent)	2005	(3.0 per cent)	13 585	(27.2 per cent)
Bit 4 ^f	37	(0.5 per cent)	33	(0.2 per cent)	45	(0.1 per cent)	77	(0.1 per cent)	1718	(3.4 per cent)
Bit 5 ^g	699	(9.2 per cent)	1922	(9.4 per cent)	9337	(21.7 per cent)	8495	(12.9 per cent)	15 206	(30.4 per cent)
Bit 6 ^h	108	(1.4 per cent)	74	(0.4 per cent)	913	(2.1 per cent)	1329	(2.0 per cent)	2161	(4.3 per cent)
Bit 7 ⁱ	1642	(21.6 per cent)	2609	(12.8 per cent)	4239	(9.8 per cent)	3426	(5.2 per cent)	5105	(10.2 per cent)
Line	O II (N)		Ne III (N)		Hd (B)		Hg (B)		Hb (N)	
Wavelength (Å)	3727		3869		4103		4342		4863	
$N_{\text{comp}}=1$	40 712	(100.0 per cent)	36 424	(100.0 per cent)	29 215	(95.2 per cent)	23 450	(91.5 per cent)	17 780	(100.0 per cent)
$N_{\text{comp}}=2$					1465	(4.8 per cent)	2159	(8.4 per cent)		
$N_{\text{comp}}=3$					4	(0.0 per cent)	18	(0.1 per cent)		
$N_{\text{comp}}=4$										
$N_{\text{comp}}=5$										
$N_{\text{comp}}=6$										
$N_{\text{comp}}=7$										
Total	40 712		36 424		30 684		25 627		17 780	
Quality										
Good ^a	21 029	(51.7 per cent)	11 498	(31.6 per cent)	8790	(28.6 per cent)	20 393	(79.6 per cent)	5123	(28.8 per cent)
Bit 0 ^b	2156	(5.3 per cent)	1028	(2.8 per cent)	3375	(11.0 per cent)	32	(0.1 per cent)	376	(2.1 per cent)
Bit 1 ^c	2184	(5.4 per cent)	1525	(4.2 per cent)	2049	(6.7 per cent)	89	(0.3 per cent)	451	(2.5 per cent)
Bit 2 ^d	9292	(22.8 per cent)	13 495	(37.0 per cent)	8711	(28.4 per cent)	2011	(7.8 per cent)	8355	(47.0 per cent)
Bit 3 ^e	9393	(23.1 per cent)	10 442	(28.7 per cent)	7994	(26.1 per cent)	374	(1.5 per cent)	4197	(23.6 per cent)
Bit 4 ^f	819	(2.0 per cent)	574	(1.6 per cent)	892	(2.9 per cent)	42	(0.2 per cent)	162	(0.9 per cent)
Bit 5 ^g	6039	(14.8 per cent)	5044	(13.8 per cent)	9380	(30.6 per cent)	1298	(5.1 per cent)	546	(3.1 per cent)
Bit 6 ^h	953	(2.3 per cent)	1064	(2.9 per cent)	1967	(6.4 per cent)	167	(0.7 per cent)	104	(0.6 per cent)
Bit 7 ⁱ	3176	(7.8 per cent)	2670	(7.3 per cent)	11 407	(37.2 per cent)	3307	(12.9 per cent)	199	(1.1 per cent)

Notes. ^a All bits set to 0.

^bBit 0: either the luminosity or its uncertainty are NaN or equal to zero.

^cBit 1: luminosity relative uncertainty > 1.5.

^dBit 2: either the FWHM or its uncertainty are NaN or equal to zero.

^eBit 3: FWHM value hits a limit in the fit.

^fBit 4: FWHM relative uncertainty > 2.

^gBit 5: either the V_{off} or its uncertainty are NaN or equal to zero.

^hBit 6: V_{off} value hits a limit in the fit.

ⁱBit 7: V_{off} uncertainty > 500 km s⁻¹.

Table C5. ...continue from Table C4.

Line Wavelength (Å)	H β (B) 4863		O II (N) 4959		O III (N) 5007		He I (B) 5876		N II (N) 6549	
$N_{\text{comp}}=1$	15 405	(86.7 per cent)	16 372	(100.0 per cent)	16 233	(100.0 per cent)	7209	(96.2 per cent)	2841	(100.0 per cent)
$N_{\text{comp}}=2$	2323	(13.1 per cent)					280	(3.7 per cent)		
$N_{\text{comp}}=3$	50	(0.3 per cent)					2	(0.0 per cent)		
$N_{\text{comp}}=4$										
$N_{\text{comp}}=5$										
$N_{\text{comp}}=6$										
$N_{\text{comp}}=7$										
Total	17 778		16 372		16 233		7491		2841	
Quality										
Good ^a	14 714	(82.8 per cent)	0	(0.0 per cent)	14 377	(88.6 per cent)	1351	(18.0 per cent)	1602	(56.4 per cent)
Bit 0 ^b	37	(0.2 per cent)	351	(2.1 per cent)	33	(0.2 per cent)	1164	(15.5 per cent)	50	(1.8 per cent)
Bit 1 ^c	42	(0.2 per cent)	322	(2.0 per cent)	30	(0.2 per cent)	407	(5.4 per cent)	64	(2.3 per cent)
Bit 2 ^d	1194	(6.7 per cent)	3384	(20.7 per cent)	500	(3.1 per cent)	3127	(41.7 per cent)	690	(24.3 per cent)
Bit 3 ^e	280	(1.6 per cent)	1838	(11.2 per cent)	1321	(8.1 per cent)	1631	(21.8 per cent)	476	(16.8 per cent)
Bit 4 ^f	16	(0.1 per cent)	111	(0.7 per cent)	11	(0.1 per cent)	236	(3.2 per cent)	6	(0.2 per cent)
Bit 5 ^g	1072	(6.0 per cent)	16 372	(100.0 per cent)	87	(0.5 per cent)	2838	(37.9 per cent)	84	(3.0 per cent)
Bit 6 ^h	76	(0.4 per cent)	0	(0.0 per cent)	32	(0.2 per cent)	858	(11.5 per cent)	20	(0.7 per cent)
Bit 7 ⁱ	1533	(8.6 per cent)	0	(0.0 per cent)	17	(0.1 per cent)	2956	(39.5 per cent)	64	(2.3 per cent)
Line Wavelength [Å]	Ha (N) 6565		Ha (B) 6565		N II (N) 6583		S II (N) 6716		S II (N) 6731	
$N_{\text{comp}}=1$	2924	(100.0 per cent)	2780	(95.1 per cent)	2843	(100.0 per cent)	2483	(100.0 per cent)	2417	(100.0 per cent)
$N_{\text{comp}}=2$			140	(4.8 per cent)						
$N_{\text{comp}}=3$			3	(0.1 per cent)						
$N_{\text{comp}}=4$										
$N_{\text{comp}}=5$										
$N_{\text{comp}}=6$										
$N_{\text{comp}}=7$										
Total	2924		2923		2843		2483		2417	
Quality										
Good ^a	1818	(62.2 per cent)	2818	(96.4 per cent)	2014	(70.8 per cent)	1136	(45.8 per cent)	1090	(45.1 per cent)
Bit 0 ^b	14	(0.5 per cent)	7	(0.2 per cent)	12	(0.4 per cent)	42	(1.7 per cent)	91	(3.8 per cent)
Bit 1 ^c	42	(1.4 per cent)	5	(0.2 per cent)	38	(1.3 per cent)	64	(2.6 per cent)	80	(3.3 per cent)
Bit 2 ^d	196	(6.7 per cent)	39	(1.3 per cent)	161	(5.7 per cent)	703	(28.3 per cent)	470	(19.4 per cent)
Bit 3 ^e	889	(30.4 per cent)	9	(0.3 per cent)	629	(22.1 per cent)	608	(24.5 per cent)	829	(34.3 per cent)
Bit 4 ^f	13	(0.4 per cent)	1	(0.0 per cent)	5	(0.2 per cent)	23	(0.9 per cent)	32	(1.3 per cent)
Bit 5 ^g	17	(0.6 per cent)	30	(1.0 per cent)	19	(0.7 per cent)	356	(14.3 per cent)	205	(8.5 per cent)
Bit 6 ^h	2	(0.1 per cent)	1	(0.0 per cent)	21	(0.7 per cent)	19	(0.8 per cent)	14	(0.6 per cent)
Bit 7 ⁱ	14	(0.5 per cent)	44	(1.5 per cent)	24	(0.8 per cent)	55	(2.2 per cent)	51	(2.1 per cent)

Notes. ^aAll bits set to 0.

^bBit 0: either the luminosity or its uncertainty are NaN or equal to zero.

^cBit 1: luminosity relative uncertainty > 1.5.

^dBit 2: either the FWHM or its uncertainty are NaN or equal to zero.

^eBit 3: FWHM value hits a limit in the fit.

^fBit 4: FWHM relative uncertainty > 2.

^gBit 5: either the V_{off} or its uncertainty are NaN or equal to zero.

^hBit 6: V_{off} value hits a limit in the fit.

ⁱBit 7: V_{off} uncertainty > 500 km s⁻¹.

This paper has been typeset from a $\text{\LaTeX}$ file prepared by the author.