Spectroscopic QUasar Extractor and redshift (z) Estimator

SQUEzE II: Universality of the results

Ignasi Pérez-Rañofs, Matthew M. Pieri

AR Marseilles Univ, CNRS, CNES, LAM, Marseille, France

ABSTRACT

In this paper we study the universality of the results of SQUEzE, a software package to classify quasar spectra and estimate their redshifts. The code is presented in Pérez-Rañofs et al. (2019). We test the results against changes on signal-to-noise, spectral resolution, wavelength coverage, and quasar brightness. We find that SQUEzE levels of performance are stable to spectra that are 4 times noisier than our standard test sample, BOSS. We also find that the performance remains unchanged if pixels of width 25˚A are considered, and decreases by ~ 2% for pixels of width 100˚A. We see no effect when changing the quasar brightness, and we establish that the blue part (up to 7000˚A) of the spectra is sufficient for the classification. Finally, we study the cases of WEAVE-QSO and DESI spectra, and J-PAS pseudo-spectra, and conclude that SQUEzE will perform similarly on them as it did on BOSS.

Keywords: cosmology: observations - quasar: emission lines - quasar: absorption lines

1 INTRODUCTION

Current spectroscopic surveys are generating hundreds of thousands of spectra of objects targeted as quasars. These spectra need to be inspected to determine whether the observed object is indeed a quasar, and to estimate their redshift when it is the case. One of such surveys is the Baryon Oscillation Spectroscopic Survey (BOSS; Dawson et al. 2013), part of the Sloan Digital Sky Survey - III survey [Eisenstein et al. (2011), with 546,856 spectra targeted as quasars that were visually inspected (Paris et al. 2017). In light of next generation of surveys such as WEAVE-QSO (Pieri et al. 2016) as part of WEAVE, Dalton et al. (2016), DESI (DESI Collaboration 2016), Euclid (Laureijs et al. 2010) and J-PAS (Benitez et al. 2014), it has become clear the community will have too many (pseudo-)spectra to process for survey-wide visual inspection to be viable.

In Pérez-Rañofs et al. (2019), hereafter Paper I, we have presented a code, SQUEzE, to automatically inspect all these spectra. SQUEzE works by measuring the presence and relative strength of potential quasar emission lines in coarse bands of spectrum using a series of metrics. The performance of the algorithm is tested using BOSS data and yields a purity of $97.40 \pm 0.47\%$ (99.59 $\pm 0.06\%$ for quasars with $z \geq 2.1$) and a completeness of $97.46 \pm 0.33\%$ (98.81 $\pm 0.13\%$ for quasars with $z \geq 2.1$) when a confidence threshold of $p_{\text{min}} = 0.32$ is used. Other codes (e.g. Quasar-Net; Busca & Balland 2018 or RedRock; an unpublished DESI code that develops the methods used by the BOSS pipeline; see Bolton et al. (2012) appear to have roughly similar performance to SQUEzE (a detailed comparison between the performance of SQUEzE and these codes will be given in follow-up papers in the series), but they have access to the entire spectra. SQUEzE, on the other hand, uses only high-level metrics. This is important because all the performance tests were made based on BOSS data (the only large enough visually inspected dataset available), but the codes will be applied to other surveys with spectra with different properties: resolution, signal-to-noise, wavelength coverage, and magnitude limit.

SQUEzE's potential for versatility is self-evident given its weak requirements on the data. This versatility is a key element of SQUEzE, as the main goal of these classifiers is to operate in unknown data coming from surveys other than BOSS (which has already finished). We test the limits of this versatility to extremes of spectroscopic data to extremely low resolution (down to pixels of 100˚A in width, compatible with filter separation in pseudo spectra obtained from many narrow band photometric surveys), and changes in quasar properties with brightness.

We therefore estimate the performance of the algorithms when applied to datasets with different properties, and determine the conditions where the trained algorithm is no longer applicable. Using only high-level metrics, SQUEzE is resistant to the changes mentioned above. The object of this paper is to determine the extent to which these changes affect the performance of SQUEzE. We will first address

---

1 Publicly available at https://github.com/iprafols/SQUEzE
how performance is affected by changes related to the usage of different instrumentation: changes in the signal-to-noise, changes in spectral binning (and effectively resolution), and changes in wavelength coverage. We also explore the impact on performance when studying fainter samples.

As in Paper I, in this paper we assume that the visual inspection is always correct. Based on a preliminary visual re-inspection of a subset of the apparent failures, we believe that the visual inspection catalogue suffers from low levels of impurity and incompleteness. Therefore a more detailed study of the validity of this statement will be addressed in a follow-up paper on the series. The datasets used here are the same used in Paper I and are described in section 3 of Paper I, but we summarize its properties here. We use 8 independent pairs of train and validation samples. Each of the samples consists of 64 plates of BOSS (~6,800 quasars and ~11,520 contaminants), and all 16 samples are independent from each other. The spectra in these samples are modified for the different tests as explained in the corresponding sections.

This paper is organised as follows. We start by testing the changes in signal-to-noise in section 2. Then we move to testing changes in spectral binning in section 3 and changes in wavelength coverage in section 4. Tests on the effect in changes on quasar properties are performed in section 5. Finally, we will discuss the results in the context of current and future surveys, and present our conclusions in section 6.

2 PERFORMANCE VS SIGNAL-TO-NOISE

As stated above, we want to estimate the performance of SQUEzE when applied to data sets of different properties. The first property we change is the signal-to-noise. We modify the signal-to-noise in BOSS data by adding Gaussian noise to the spectra. The noise is added at the pixel level as

\[ f_i' = f_i + (N_{\text{noise}} - 1) \sigma_i G(0, 1), \]

\[ \sigma_i' = \sigma_i \sqrt{N_{\text{noise}}}, \]

where \( f_i \) and \( \sigma_i \) are the flux and standard deviation of a pixel in a given pixel, \( f_i' \) and \( \sigma_i' \) are the modified fluxes and standard deviations, \( G(0, 1) \) is a random number drawn from a Gaussian with mean 0 and standard deviation 1.

The case \( N_{\text{noise}} = 1 \) corresponds to the original data, and we explore the cases \( N_{\text{noise}} = 2, 3, 4 \), corresponding to having 2, 3, and 4 times the original noise. We name these cases noise2, noise3, and noise4. Figure 1 illustrates the noise addition in a randomly selected spectrum and the behaviour of the peak finder in each of the cases. We can see that in the original spectrum we successfully identify the Lyβ, Lyα, SiIV, CIV, and CIII emission lines (see the restframe wavelength for these lines in Table 1). When we increase the noise we maintain the detection of Lyα and CIV, whereas we lose the weaker Lyβ and SiIV lines. Also, we note that more noise peaks are found in noiser data. As expected, as noise increases the peak finder fails to detect the less prominent peaks. However, because spectra are smoothed before looking for peaks, real peaks are still found even with higher noise levels.

We then test the performance of SQUEzE on these noise-augmented spectra. We run SQUEzE twice, first retraining the models on the modified data, and then using the original training. Figure 2 shows the results of this test taking values of \( p_{\text{min}} \) such that purity is approximately equal to completeness for the entire sample (also the standard choice in Paper I). In the top panel, we see that the performance decreases by ~1% for case noise2, by ~3–4% for case noise3, and by ~7% for case noise4. While the decrease is expected, we note that SQUEzE is obtaining purity and completeness of around 91% even when the noise variance is 4 times that of the original sample. For Lyα quasars (with \( z \geq 2.1 \)) the results are even more stable: purity decreases by only ~1% for case noise4, while completeness decreases by ~3%. In the bottom panel we show the difference in performance when using the original or the retrained models, where we see that there is essentially no need to retrain if small changes on the noise levels are expected. If the sample is significantly noisier, as in the case noise4, then retraining the models minimizes losses. We conclude that SQUEzE is mildly affected by the noise level in the spectra.

Let us analyze in more detail the case noise4. We have seen in Figure 2 that there is a small loss changes in the performance is detected when the noise variance is increased by a factor of 4, and that retraining the model minimizes this loss. This is the case when choosing a value of \( p_{\text{min}} \) such that purity and completeness are similar. If we analyze the dependence of this statement on the chosen value of \( p_{\text{min}} \) (Figure 3) we again see that retraining the model maximizes the performance.

3 PERFORMANCE VS BINNING/RESOLUTION

Now that we have seen that SQUEzE is insensitive to changes in the signal-to-noise, we shift our attention to changes in the resolution. Now we modify the original data by rebinning the data into wider bins. The rebinned flux is computed by averaging the fluxes of the original bins that fall into the new bins. The rebinned errors are computed by standard error propagation. The new bins are created so that there is a bin centered at 4,000Å and have widths 3.125, 6.25, 12.5, 25, 50, and 100Å. For reference, the size of the original pixels is ~1Å. Figure 4 illustrates this rebinning and the behaviour of the peak finder on rebinned data. Again, we can see that in the original spectrum we successfully identify the Lyβ, Lyα, SiIV, CIV, and CIII emission lines (see the restframe wavelength for these lines in table: 0.000–000 (0000))

| line      | wavelength [Å] |
|-----------|----------------|
| Lyβ      | 1033.03        |
| Lyα      | 1215.67        |
| SiIV     | 1396.76        |
| CIV      | 1549.06        |
| CIII     | 1908.73        |
| NeIV     | 2423.83        |

Table 1. Summary of the properties of the selected lines. The columns show the line’s name, the nominal wavelength of the line, and the starting and ending wavelengths for the emission, blue, and red windows. These windows are used to compute the line ratio, the line contrast ratio and the line continuum slope, as explained in the text. All wavelengths are given in Å at the restframe.
Once we have rebinned the spectra we run SQUEzE on them to assess its performance. As before we run SQUEzE twice, both retraining the model on the modified data, and using the original training. We show the results of this test in Figure 6 where again we choose values of $p_{\text{min}}$ such that purity and completeness are similar. We see that there is no change in the performance for bins of sizes up to 25Å. Up to bins of this scale, retraining the model does not improve the performance. The performance decreases to a small but significant degree for bin sizes of 50 and 100Å: ~2% even for bin size of 100Å. Uniquely among our rebinning tests, 100Å bins using only the original training leads to a significant impact on performance compared to retraining: an additional loss of ~2%. For Ly$\alpha$ quasars the decrease in the performance is less prominent: purity is not modified and completeness drops by ≤1.5%. We conclude that SQUEzE is resistant to changes in the resolution, provided that the pixel size is smaller than 25Å. When binning further to 100Å, retraining based on equivalent data helps alleviate the decrease in performance, and for Ly$\alpha$ quasars this retraining appears entirely successful.

We now analyze in more detail the case rebin 100. As already seen in Figure 5, the performance of SQUEzE is decreased in this case study, and the decrease is significantly worse if the original training is used. In Figure 6 we show the validity of this statement as a function of $p_{\text{min}}$. We see that retraining improves the performance of SQUEzE, especially in the completeness at all redshifts, and for high values of $p_{\text{min}}$. We note an increase in purity is observed at $z ≥ 2.1$ for the retrained model at low values of $p_{\text{min}}$.

A possible (naive) explanation for the decrease in the performance of SQUEzE in the case rebin100 would be that the rebinning is moving the center of the emission peaks, and therefore the estimated quasar redshift. In this scenario, quasar redshift would be larger than the tolerance redshift, $Δz = 0.15$, discarding some quasars that are correctly identified. If this explanation were correct, we would see quasar contaminants very close to the $z_{\text{true}} = z_{\text{true}}$ line in the line confusion plot (Figure 7). Because we do not see such a thing, we rule out this explanation. Thus, it appears that at this level of binning the calculation of the metrics themselves becomes compromised since the bins widths approach the size of bands used in their calculations.

### 4 PERFORMANCE WITH LIMITED WAVELENGTH COVERAGE COVERAGE

We move now to evaluate the performance of SQUEzE as a function of wavelength coverage. We explore this by returning to the original data and removing 1/4 and 1/2 of the total coverage. We analyse five cases: red1, red2, blue1, blue2, and mid. Their wavelength coverage is specified in Table 3.

Figure 8 summarizes the result of this exercise. SQUEzE performance is strongly reduced for samples mid, red1, and particularly red2. The samples blue1 and blue2, on the other hand, present a much milder decrease in performance. This suggests that the blue end of the spectra are driving the classification. To better understand this behaviour we analyze the line confusion plots for samples blue1 and red1 (Figures 9 and 10 respectively). Note that we select a $p_{\text{min}}$ value such...
Figure 2. Top panels: Purity and completeness as a function of the noise level (see equation 1). For each sample we select a $p_{\text{min}}$ such that purity and completeness are similar for the entire sample and the right panels show the performance for these samples. The left panels show the performance limiting ourselves to quasars with $z \geq 2.1$. Points are horizontally shifted to avoid overlap. Top panels: Retraining and testing on the same noise-augmented data. Bottom panels: difference in the performance between using the original or the retrained models. Negative (positive) values indicate that the retrained (original) models are better. $N = 1$ correspond to the original data in Paper I.

Figure 3. The dependence of the noise4 results using the original or the retrained models for varying probability thresholds. The plot shows the purity (top panels) and the completeness (bottom panels) measured on the test sample with the noise variance increased by a factor of 4. Red and blue lines show the results when the classifier is trained with the original data and with the noise-augmented data, respectively. Left panels shows the results for $z_{\text{try}} \geq 2.1$, whilst right panels show the results at all redshifts. Dashed vertical lines show the $p_{\text{min}}$ used in figure 2.
Figure 4. Example of the performance of the peak finder with spectra with different resolution.

Figure 5. Similar to Figure 2 but as a function of the resolution. Pixel size of 1 Å correspond to the original data.
Figure 6. The dependence of the rebin100 results using the original or the retrained models for varying probability thresholds. The plot shows the purity (top panels) and the completeness (bottom panels) measured on the test sample with the noise variance increased by a factor of 4. Red and blue lines show the results when the classifier is trained with the original data and with the rebinned data, respectively. Left panels show the results for $z_{\text{true}} \geq 2.1$, whilst right panels show the results at all redshifts. Dashed vertical lines show the $p_{\text{min}}$ used in figure 5.

Figure 7. Line confusion plot for the case study rebin100, where the model has been retrained. For all the objects in the catalogue, $z_{\text{true}}$ against $z_{\text{true}}$. In this plot, green circles correspond to correct classifications, yellow circles to stellar contaminants, red circles to galactic contaminants, and blue circles to quasar contaminants. Note that this plot signals out the false positives.

Table 3. Wavelength coverage of the original data and the five case studies.

| Sample | Wavelength coverage [Å] |
|--------|-------------------------|
| orig   | 3,600-10,400            |
| red1   | 5,300-10,400            |
| red2   | 7,000-10,400            |
| blue1  | 3,600-8,700             |
| blue2  | 3,600-7,000             |
| mid    | 5,300-8,700             |

that purity and completeness are similar for each of the samples and are, therefore, different. We see that the decrease on purity and completeness for sample red1 is mostly due to line confusion. Loosing the blue part of the spectrum is making SQUEzE misidentify the peaks. We note that the quasar emission lines are closer together in the blue part of the spectrum, as seen in figure 1 of Pérez-Rafols et al. (2019).
original training is expected to perform worse than the retrained model.

5 PERFORMANCE VS QUASAR BRIGHTNESS

Thus far we have presented the impact on performance when data from surveys other than BOSS are used. In particular, we have focused only on differences whose origins arise due to changes in instrumentation. Important performance modifications may also arise due to changes in the quasar target sample. We explore this point here by testing the impact of quasar brightness over the dynamic range provided by the BOSS sample.

One of the major assumptions of SQUEzE is that quasars are self-similar, but in reality this is just an approximation: we know that there is a quasar-to-quasar variation. This variation arises associated a number of things: presence of broad absorption lines, damped Lyman-α absorbers, difference in continuum slope(s) and differences emission line strengths. All this variation exists in our test samples, but if the survey on which SQUEzE is implemented differs significantly in its realization of these various effects, SQUEzE may not perform as we have described. It is clear that one way future surveys will differ is in their luminosity distributions, but targeting to a fainter limiting magnitude. An example of such a change is the well-known Baldwin effect (Baldwin 1977), which states that the relative strength of the emission lines depends on the luminosity.

While we cannot yet test SQUEzE on significant numbers of quasars fainter than the BOSS limit (r ≤ 22.2), we can test the sensitivity to quasar faintness by taking brightness dependent sample subsets. To this end, we split each of the samples into two bright/faint samples of roughly equivalent signal-to-noise by obtaining approximately equal values for the ensemble quantity

\[ \sum \left( \frac{r_i}{\delta r_i} \right)^2, \]  

(3)

where \( r_i \) and \( \delta r_i \) are the r-band magnitude and its error \( \delta \) of a spectrum \( i \) (quasar or contaminant alike) belonging to the subsample.

The resulting boundary between bright and faint samples is \( r \sim 16 \). We apply models trained using the bright samples (bright models) to both the bright and the faint validation samples. Similarly we apply models trained using the faint samples (faint models) to both bright and faint validation samples.

Results of this exercise are shown in Figure 11. We observe that SQUEzE performance on \( z > 2.1 \) quasars is not significantly affected by these sample changes. It is particularly striking that SQUEzE performs well on faint \( z > 2.1 \) quasars, including quasars as faint as \( r \sim 22 \), even when it is trained using quasars brighter than \( r \sim 16 \).

This does not remain the case when all redshifts are included. It would appear that the performance for faint quasars does indeed benefit from retraining on similar quasars but only at the ~ 2–3% level and that it is driven by a dependence on the weaker emission lines with wavelengths long-ward of Lyman-α.

6 DISCUSSION AND CONCLUSIONS

In this paper we have explored the universality of SQUEzE results (presented in Paper I). In particular, we have addressed the effect of changing the signal-to-noise, the spectral binning (and resolution), the wavelength coverage and the brightness of the sample. After all these tests we are confident that the performance of SQUEzE is largely survey-independent. However, for optimal performance we do recommend altering BOSS spectra to resemble the characteristics of the survey to be analysed and retraining the model on the modified data. If more precise and reliable tests are required a more realistic contamination sample matched to the details of the given survey must be explored (such a sample could be obtained, for example, in the survey validation phase). We do not recommend retraining using synthetic spectra, unless full modelling of the contaminants is also taken into account.

All the tests performed here assume that the visual inspection is always correct, but this is not necessarily the case. We will explore this in more detail in a subsequent paper of the series. However, we remark that the tests are fair since what we analyse is the difference in the results on the original and modified spectra, and whatever misclassifications are present in truth table for the original samples are also present in the modified samples.

A study of the universality of the results is important since we expect SQUEzE to be used in other surveys whose characteristics will differ from those in BOSS. In the near future, the two most relevant surveys are WEAVE-QSO and DESI. In both cases the minimum signal-to-noise to full depth for quasars with \( z > 2.1 \) will be similar to that of BOSS and therefore the results shown in Paper I are already a reasonable guide to quasar catalogue making performance with regards to noise. However, the quasar samples may themselves differ since these surveys reach a limiting magnitude of \( r \sim 23 \). While we do not assess this limit we do test the importance of quasar brightness. We find that within the BOSS magnitude distribution, SQUEzE performs satisfactorily upon faint \( z \geq 2.1 \) quasars even when training using bright \( z \geq 2.1 \) quasars. When including \( z < 2.1 \) quasars retraining on quasars with representative magnitudes improves performance by ~ 2–3%. Both DESI and WEAVE-QSO provide higher spectral resolution than BOSS but our binning tests show that a factor of ~ 2 or ~ 3 respectively will have no impact on SQUEzE performance.

For both DESI and WEAVE-QSO, significant challenges are faced with regards to building the survey. In the case of DESI, only \( z > 2.1 \) quasars will be observed to full depth in four layers (or passes) over the survey footprint. Quasars must be efficiently identified (and assigned approximately redshifts) based on a single layer (or a quarter of the exposure time) to either determine whether to continue in subsequent observing layers or to refine the redshift with the data as-is. This is equivalent to our noised case (see tests in section 2). We find that SQUEzE will be able to identify
$z \geq 2.1$ quasars on the first pass to high purity but that completeness may be reduced by a few percent (depending on the choice of probability threshold). Including quasars with $z < 2.1$ presents more of a challenge with a larger impact on both purity and completeness (of approximately 7%) that cannot be recovered trivially by tuning the probability threshold.

In the case of WEAVE-QSO, again SQUEzE is expected to perform well for catalogue production and as a redshift prior for additional redshift refinement step. The main challenge of quasar identification for WEAVE-QSO is in its requirement of highly complete and highly pure quasar target sample provided by the survey J-PAS. J-PAS is a narrow-band photometric survey with 56 filters such that J-PAS data can be considered as pseudo-spectra. These pseudo-spectra have an effective binning of $\sim 100\,\AA$. Our results with aggressive binning show SQUEzE performance even at the level of $100\,\AA$ bins. The purity of $z > 2.1$ seems to
be largely unaffected and the completeness show only a 1% loss. The WEAVE-QSO is entirely focussed on Lyman-α forest quasars, so this is sufficient for its requirement but it is noteworthy that the performance at all redshifts remains strong.

An additional challenge for WEAVE-QSO is reflected in the tight schedule of J-PAS targeting and WEAVE observing. As a result it is likely that only some of the filters will be observed in time for propagation to WEAVE-QSO fiber assignment. In this regard SQUEzE performance tests with limited wavelength coverage could be critical. The success of the blue2 test (using only the blue half of the optical range) indicates a potential way forward for the planning for J-PAS and WEAVE-QSO scheduling combined with SQUEzE. It might be noted, however, that these are isolated tests and true quasar identification in J-PAS will involve a combination of increased noise (since the limiting magnitude will be 1 magnitude fainter than our sample), ∼100Å binned and limited wavelengths coverage. As a result realistic J-PAS mocks are needed to address these challenges simultaneously.

Other surveys such as Euclid or WEAVE-LOFAR (another survey as part of WEAVE) may also benefit from the use of SQUEzE, but their specific requirements have not yet been explored.

To conclude we find that SQUEzE is not only effective on BOSS-like spectra of quasar targets (as shown in Paper I), but is also independently resistant to increases in noise variance by a factor 4, increases in spectral binning by a factor 100, halving of the wavelength coverage, and changes in quasar brightness of 6 magnitudes.

ACKNOWLEDGMENTS

This work was partly supported by the A*MIDEX project (ANR-11-IDEX-0001-02) funded by the “Investissements d’Avenir” French Government program, managed by the French National Research Agency (ANR), and by ANR under contract ANR-14-ACHN-0021.

References

Baldwin, J. A. 1977, ApJ, 214, 679
Benitez, N., Dupke, R., Moles, M., et al. 2014, arXiv e-prints, arXiv:1403.5237
Bolton, A. S., Schlegel, D. J., Aubourg, ´E., et al. 2012, AJ, 144, 144
Busca, N., & Ballard, C. 2018, arXiv e-prints, arXiv:1808.09955
Dalton, G., Trager, S., Abrams, D. C., et al. 2016, in Proc. SPIE, Vol. 9908, Ground-based and Airborne Instrumentation for Astronomy VI, 99081G
Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. 2013, AJ, 145, 10
DESI Collaboration. 2016, arXiv e-prints, arXiv:1611.00036
Eisenstein, D. J., Weinberg, D. H., Agol, E., et al. 2011, AJ, 142, 72
Laureijs, R. J., Duvet, L., Escudero Sanz, I., et al. 2010, in Proc. SPIE, Vol. 7731, Space Telescopes and Instrumentation 2010: Optical, Infrared, and Millimeter Wave, 77311H
Pérès-Rafols, I., Pieri, M. M., Blomqvist, M., Morrison, S., & Som, D. 2019, arXiv e-prints, arXiv:1903.00032
Pieri, M. M., Bonoli, S., Chaves-Montero, J., et al. 2016, in SF2A-2016: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics, 259–266

MNRAS 000, 000–000 (0000)