Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **498,** 5227–5239 (2020)
Advance Access publication 2020 September 18

doi:10.1093/mnras/staa2826

Automated measurement of quasar redshift with a Gaussian process

Leah Fauber,[1] Ming-Feng Ho,[1] Simeon Bird,[1]★ Christian R. Shelton,[1] Roman Garnett[2] and Ishita Korde[1]

[1]*University of California Riverside, Department of Physics and Astronomy and Department of Computer Science, Riverside, CA 90703, USA*
[2]*Washington University in St. Louis, Department of Computer Science and Engineering, One Brookings Drive, St. Louis, MO 63130, USA*

Accepted 2020 September 11. Received 2020 September 9; in original form 2020 June 22

ABSTRACT

We develop an automated technique to measure quasar redshifts in the Baryon Oscillation Spectroscopic Survey of the Sloan Digital Sky Survey (SDSS). Our technique is an extension of an earlier Gaussian process method for detecting damped Lyman α absorbers (DLAs) in quasar spectra with known redshifts. We apply this technique to a subsample of SDSS DR12 with BAL quasars removed and redshift larger than 2.15. We show that we are broadly competitive to existing quasar redshift estimators, disagreeing with the PCA redshift by more than 0.5 in only 0.38 per cent of spectra. Our method produces a probabilistic density function for the quasar redshift, allowing quasar redshift uncertainty to be propagated to downstream users. We apply this method to detecting DLAs, accounting in a Bayesian fashion for redshift uncertainty. Compared to our earlier method with a known quasar redshift, we have a moderate decrease in our ability to detect DLAs, predominantly in the noisiest spectra. The area under curve drops from 0.96 to 0.91. Our code is publicly available.

Key words: methods: statistical – quasars: absorption lines – quasars: emission lines – quasars: general.

1 INTRODUCTION

Estimating redshifts using spectroscopy is a well-explored technique in astronomy. Spectroscopy uses the presence of lines at known emission wavelengths to estimate the redshift of an object. While quasi-stellar objects (QSOs, or quasars) contain multiple strong emission lines, the presence of quasar outflows mean that these lines often have an intrinsic Doppler shift from their rest positions, leading to hard to quantify redshift errors (Gaskell 1982; Shen et al. 2016). The Sloan Digital Sky Survey (SDSS; Eisenstein et al. 2011; Dawson et al. 2013; Alam et al. 2015) presents a further challenge due to the low signal to noise of many of the spectra. Redshift estimation in Data Release 14 (DR14Q) is done using four different techniques. These include principal component analysis (PCA) using DR5 as a training sample (Hewett & Wild 2010; Schneider et al. 2010), automated fitting to the Mg II emission line, and a partial visual inspection survey (Pâris et al. 2018). Techniques differ, on average, by around 100 km s^{-1}, with a velocity dispersion of $\sim$500 km s^{-1}. Furthermore, they fail to converge for about 0.5 per cent of objects. Estimation of quasar redshift, $z_{\rm QSO}$, must be accurate to achieve the scientific goals of spectroscopic surveys. Systematic and statistical errors in redshift estimation reduce the strength of the Baryon Acoustic Oscillation (BAO) signal (Dawson et al. 2016).

Each new generation of spectroscopic survey roughly doubles the number of quasar spectra, such that DR14Q contains 1.8×10^5 quasars with Lyman α absorption in the Baryon Oscillation Spectroscopic Survey (BOSS; Pâris et al. 2017). The next-generation Dark Energy Spectroscopic Instrument (DESI) will ultimately contain 7×10^5 Lyman α quasars (DESI Collaboration 2016). Algorithmic inspection of quasar spectra, already essential, will become yet more necessary to keep pace with data collection.

We estimate quasar redshifts using a Gaussian process (GP) model for quasar spectra. Compared to existing redshift estimation techniques, our model is conceptually most similar to PCA redshifts, although we improve on them by explicitly accounting for noise in the spectrum. All emission lines in the redshift range are fit simultaneously. Our model uses the existing catalogue as a prior to constrain the expected offsets of each line from the intrinsic emission redshift. In principle, we are also able to learn correlations between emission line width and velocity offset (Mason, Brotherton & Myers 2017).

We build on the work of Garnett et al. (2017), Bird, Garnett & Ho (2017), and Ho, Bird & Garnett (2020). Garnett et al. (2017) built a GP model for quasar spectra and combined it with an analytic Voigt profile to find damped Lyman α absorbers (DLAs), strong neutral hydrogen absorption lines corresponding to the gas surrounding high-redshift dwarf galaxies (Wolfe et al. 1986; Prochaska & Wolfe 1997; Haehnelt, Steinmetz & Rauch 1998; Bird et al. 2014). We extend the emission model implicit in that work to the whole quasar spectrum between 3000 and 910 Å. We then try to use all information about the shape and properties of the quasar to estimate the quasar redshift. In practice redshift estimation in our model is driven by the fit to well-known emission peaks, especially Mg II, C III, C IV, and Lyman α. We train the model using the SDSS pipeline quasar redshift estimate, and use the trained model to estimate the redshift of quasars outside the training set. To verify our method, we check our derived quasar redshifts against the other redshift estimates included in the SDSS catalogue and show that they are competitive to other techniques.

We also provide a modified DLA catalogue for SDSS DR12 to demonstrate that we can detect DLAs while marginalizing out redshift uncertainty. To validate the results, we compare them to catalogues from the template fitting code of Noterdaeme et al. (2012),

★ E-mail: sbird@ucr.edu

the SDSS visual inspection survey and the neural network based model of Parks et al. (2018). We require three separate catalogues in order to generate and thus compare to a 'best two of three' catalogue to serve as ground truth. We emphasize that we use visual inspection, a non-automated technique which will not be available for future data releases, only for validation.

Section 2 defines our overall emission model, Section 3 describes our redshift estimation, and Section 4 describes DLA finding. We summarize the main results from Garnett et al. (2017), on which our model is heavily based and point explicitly to changes. Our model for DLAs includes most of the updates presented in Ho et al. (2020), but for computational reasons finds only one DLA per spectrum and does not include the sub-DLA model. Section 5 summarizes our training set. Our results are presented in Section 6 and we conclude in Section 7. Our redshift estimation code is available on GITHUB at https://github.com/sbird/gp_qso_redshift. Our DLA model with redshift estimation may be found at https://github.com/sbird/gp_dla_detection/tree/zqsos2.

2 A GAUSSIAN PROCESS MODEL FOR QSOS

Quasar emission spectra are complex functions which do not have a known closed parametric form. Our method builds a model for the expected shape of a quasar emission spectrum $f(\lambda)$. We use a Gaussian process (GP; Williams & Rasmussen 2006), a non-parametric framework able to model complex continuous functions. Importantly, a Gaussian process can describe how variations in the observed spectra are correlated as a function of wavelength. The learned model will naturally include information describing the presence of emission lines. The training set for our model is SDSS DR12 with broad absorption line (BAL) quasars removed and $z_{\rm VI} < 2.15$. We use SDSS visual inspection redshift estimates during training. However, the trained model is applicable to larger, unlabelled, data sets. After training, the learned model is used to evaluate the likelihood function of each quasar spectrum as a function of redshift. Our point redshift estimate is located at the maximum a posteriori value of the likelihood function and the redshift uncertainty is given by 95 per cent confidence intervals.

A GP is a generalization of the Gaussian distribution which describes random functions, rather than random vectors. Naively, we can think of a GP as a Gaussian distribution extended over an infinite number of dimensions. It is described by a mean *function*, $\mu(\lambda)$ and a covariance *function*, $K(\lambda_1, \lambda_2)$. The mean describes the average value of a draw (of a function) from the GP. The covariance describes the correlations between any two points on the function, $f(\lambda_1)$ and $f(\lambda_2)$. If a fixed set of regressors, for example $\lambda_1, \lambda_2, \ldots, \lambda_m$, is selected, the random function evaluated at these values generates a set of (dependent) random variables: $f(\lambda_1), f(\lambda_2), \ldots, f(\lambda_m)$. In a GP, these random variables are jointly Gaussian. Their means are just the application of μ to the independent values, and the covariance matrix is similarly constructed: $E[f(\lambda_i)] = \mu(\lambda_i)$ and covar$[f(\lambda_i), f(\lambda_j)] = K(\lambda_i, \lambda_j)$.

There are no off-the-shelf Gaussian process covariance functions able to model the complex shape of a quasar. We thus learn a covariance function from the training data. Our model assumes that the emission spectrum from a QSO (in its rest frame), y, is drawn (independently from $z_{\rm QSO}$) from a Gaussian process with mean function μ and a covariance function K, which we denote as

$$p(y) = \mathcal{N}(y; \mu, K). \tag{1}$$

We choose to build our GP model at the rest-frame $\mathrm{rs}_{z_{\rm QSO}}(\lambda_{\rm OBS})$. We therefore can capture the covariance between different emission lines from different quasars by setting them on to the same rest-wavelength pixels. The relationship between the rest frame and observed frame is

$$\mathrm{rs}_{z_{\rm QSO}}(\lambda_{\rm OBS}) = \frac{1}{1+z_{\rm QSO}}\lambda_{\rm OBS}. \tag{2}$$

The Gaussian process describing the QSO spectrum can be transformed into the observed frame, and remains a Gaussian process. Letting $\tilde{y}$ be the emission spectrum in the observed frame,

$$p(\tilde{y}) = \mathcal{N}\left(\tilde{y}; \mu \circ \mathrm{rs}_{z_{\rm QSO}}, K \circ \mathrm{rs}_{z_{\rm QSO}}\right),$$
$$(\mu \circ \mathrm{rs}_{z_{\rm QSO}})(\lambda_{\rm OBS}) = \mu(\mathrm{rs}_{z_{\rm QSO}}(\lambda_{\rm OBS})),$$
$$(K \circ \mathrm{rs}_{z_{\rm QSO}})(\lambda_1, \lambda_2) = K\left(\mathrm{rs}_{z_{\rm QSO}}(\lambda_1), \mathrm{rs}_{z_{\rm QSO}}(\lambda_2)\right).$$

The observed spectrum, x, is equal to $\tilde{y}$, but after absorption between the observer and the quasar and additive noise from the observational instrument. Calculating the scale factor $\frac{1}{1+z_{\rm QSO}}$ requires knowledge of the quasar redshift. Let $\mathcal{D} = (\lambda_{\rm OBS}, x)$ be a set of quasar observations in the observed frame, where $\lambda_{\rm OBS}$ is the set of wavelengths in the observed frame, and x is the set of observed flux. We learn our GP model D^0 at $\mathrm{rs}_{z_{\rm QSO}}(\lambda_{\rm OBS})$ using a training set of observations with known quasar redshifts, $\mathcal{D} = (\mathrm{rs}_{z_{\rm QSO}}(\lambda_{\rm OBS}), x, z_{\rm QSO})$, where $z_{\rm QSO}$ is the redshift estimated by the SDSS pipeline. After we learn the GP model D^0, we use observations outside the training set $\mathcal{D} = (\lambda_{\rm OBS}, x)$ to validate our D^0.

We assume that absorption between the observer and the quasar and additive noise from the observational instrument are independent of each other and that both are uncorrelated between wavelength bins. The instrument noise is modelled using a Gaussian process with a zero mean function and a 'diagonal' covariance kernel. $K(\lambda_1, \lambda_2)$ is zero if λ_1 and λ_2 are not equal (or almost equal). Instrument noise is a property of the survey, and is not learned during training. If K_N is the kernel for the instrument noise, the observed spectrum, x, is also drawn from a Gaussian distribution if we condition on $z_{\rm QSO}$:

$$p(x|z_{\rm QSO}) = \mathcal{N}\left(x; \mu \circ \mathrm{rs}_{z_{\rm QSO}}, (K \circ \mathrm{rs}_{z_{\rm QSO}}) + K_N\right).$$

Section 4.1 describes neutral hydrogen absorbers in the intergalactic medium, which are treated separately. As they do not strongly affect the shape of the peaks which dominate the redshift estimation, we neglect them except when finding DLAs. We have not attempted to model BAL and have removed BAL quasars from the sample.

3 LEARNING A GP FOR REDSHIFT ESTIMATION

In this section, we describe the modelling decisions we made to extend our Gaussian process model, D^0, for quasar redshift estimation. D^0 is a lightweight GP model and may be sampled to obtain the likelihood of the quasar redshift, $z_{\rm QSO}$. The shape of this likelihood in turn produces $p(z_{\rm QSO}|x, \mathrm{D}^0)$, the posterior distribution for $z_{\rm QSO}$.

The null model D^0 contains information describing the average shape of a quasar. A minimal modification of Garnett et al. (2017) would fit this null model to different quasar redshifts. We found however, that this minimal modification does not have sufficient information to fit the quasar. We thus modify it in two important ways. First, we extend the modelled Gaussian process range to 910−3000 Å, in order to encompass more emission lines, especially Mg II. Secondly, we augment the model to explicitly model the likelihood of observations outside the modelled redshift range. There is thus some likelihood component for all observations and so probabilities are comparable for the same spectrum across multiple redshifts.

3.1 Redshift prior

In this paper, we treat $z_{\rm QSO}$ as a parameter to estimate, rather than a known value. We place a bounded uniform prior on the parameter $z_{\rm QSO}$, $p(z_{\rm QSO})$:

$$p(z_{\rm QSO}) = \mathcal{U}[z_{\rm QSOmin} - z_\epsilon, z_{\rm QSOmax} + z_\epsilon], \quad (3)$$

where $z_{\rm QSOmin}$ and $z_{\rm QSOmax}$ are the minimum and maximum quasar redshifts. For our SDSS sample they are 2.15 and 6.44, respectively. We extend the prior range by a small amount ($z_\epsilon = 3000$ km/s) on either side to ensure that no samples lie on the prior boundaries. We use a uniform prior rather than a data-driven prior to demonstrate that our method is applicable to arbitrary quasar spectra within the prior range, rather than just the SDSS data set.[1]

3.2 Extended model range

The original modelling range of D^0 ran from the rest-frame Lyman limit (910 Å) to the rest-frame Lyman α (1216 Å). We extend this range to cover much of the metal line region. In the rest frame

$$\mathrm{rs}_{z_{\rm QSO}}(\lambda_{\rm OBS}) = \lambda_{\rm rest} \in [910\,\text{Å}, 3000\,\text{Å}]. \quad (4)$$

An extension to 3000 Å allows us to include the Mg II emission line (2799 Å). Mg II is a particularly valuable emission line as it is the least affected by systemic velocity shifts (Hewett & Wild 2010; Shen et al. 2016). The pixel spacing remains the same as that of Garnett et al. (2017) with $\Delta\lambda = 0.25$ Å, giving us 8 361 pixels in our GP mean vector.

Bluewards of the Lyman limit, the occasional presence of strong absorption from a Lyman limit system introduces substantial variance into the model, so that it has little redshift constraining power. Furthermore, this region is hard to train. Only relatively rare $z_{\rm QSO} > 3.7$ quasars contain rest-frame data at $z < 910$ Å. We thus exclude the region bluewards of the Lyman limit from the modelling range of the Gaussian process.

To model the relationship between quasar flux measurements and the true QSO emission function, we have to include the correlation between emission lines K and the instrumental noise K_N. When we are only interested in estimating redshift, we do not include the model for neutral hydrogen absorption ('Lyman α absorption noise' in our earlier papers). This model affects only the continuum bluewards of the Lyman α peak, which has relatively large instrumental noise compared to the metal-line region and is thus sub-dominant when estimating redshift. We have confirmed that this approximation does not significantly affect our results, yet it reduced the training time for the model by a factor of ∼20.

3.3 Observed data outside GP range

As we do not model the entirety of the quasar spectrum, our likelihood is incomplete. We would like to evaluate the marginal likelihood of the GP to estimate $z_{\rm QSO}$. However, to ensure that we can compare posterior probabilities at different redshifts, we need to provide a likelihood function for the data not modelled by the main GP. Otherwise, as different observations fall into the model, likelihoods are evaluated on different subsets of the data. To avoid this problem, we implemented an explicit model for observed data outside the Gaussian process model boundaries. All observed data is thus accounted for in the extended model.

[1]Note that for the DLA finding problem we use a different, data-driven, prior as we integrate out $z_{\rm QSO}$ to find $\log_{10} N_{\rm H\,I}$ and $z_{\rm DLA}$.

To illustrate the need for this model, consider when emission peaks are redshifted out of the GP model range. A $z \sim 2.5$ quasar assumed to be at $z = 5$ will have the emission corresponding to the Lyman α emission peak at 1216 Å incorrectly appear at 700 Å, outside the modelled rest frame. As the peak is now outside the rest frame, D^0 applies no penalty for not predicting the emission peak and may incorrectly prefer a high redshift.

Our explicit extra model assumes that the emission spectrum in the rest-frame bluewards of 910 Å are drawn independently and identically from a Gaussian distribution with a constant variance. We make the same assumption for those emission spectrum values redwards of the GP model's range. These 'out-of-GP' emission fluxes are subject to the same instrument noise and absorption as the rest of the spectrum, after being transformed to the observer frame. However, they have no correlations with each other or with the flux modelled by the GP in 910–3000 Å.

The mean and standard deviations of these two Gaussian distributions are optimized for during training. We define $\mu_{\rm red}$ and $\sigma_{\rm red}$ to be the mean and standard deviations of the 'out-of-GP' model for the redward end. If $\sigma_{\rm red}$ is known, the maximum likelihood estimate for $\mu_{\rm red}$ can be computed in closed form:

$$\mu_{\rm red} = \frac{\sum_i \rho_i x_i}{\sum_i \rho_i}, \quad (5)$$

where

$$\rho_i = \frac{1}{\sigma_i^2 + \sigma_{\rm red}^2}. \quad (6)$$

Here, i ranges over observations in the training set that fall redwards of the Gaussian process model and x_i is the observed flux (recall that the training data have known $z_{\rm QSO}$ values). σ_i denotes the standard deviation of the instrumental noise for observation i. Thus, each observation, x_i, is drawn independently from a normal distribution with mean $\mu_{\rm red}$ and variance $\sigma_{\rm red}^2 + \sigma_i^2$.

To find $\sigma_{\rm red}$, we conduct a line search to find the maximum likelihood, using the above substitution for $\mu_{\rm red}$ in terms of $\sigma_{\rm red}$ in the likelihood. The resulting function (ignoring constants) to be optimized is

$$\log \mathcal{L} = \sum_i \rho_i \, (x_i - \mu_{\rm red})^2 \, - \log \rho_i, \quad (7)$$

where ρ_i and $\mu_{\rm red}$ both depend on $\sigma_{\rm red}$, the quantity to be tuned. Empirically, this likelihood is concave and easy to maximize. The fitting procedure for the blueward end model is identical, but on a different set of fluxes.

3.4 Quasar normalization

The observed magnitude of a quasar depends on its luminosity distance and the properties of the black hole. To allow a single GP model to describe the observed flux x, we normalize the flux measurements. Garnett et al. (2017) chose to normalize at an absorption free region between 1310 and 1325 Å in the rest frame. Here, we change the normalization range to 1176−1256 Å for building D^0, normalizing all spectra at the same Lyman α peak amplitude.

We choose to normalize the amplitude of the quasar spectrum to the Lyman α peak region, 1216 ± 40 Å;. We found empirically that this produced the most accurate quasar redshift estimation during our validation experiments. The position of the Lyman α peak is highly variable, which may at first make it seem a poor choice for normalization. We emphasize however that we only use the peak height, and not the peak position, to normalize the overall quasar

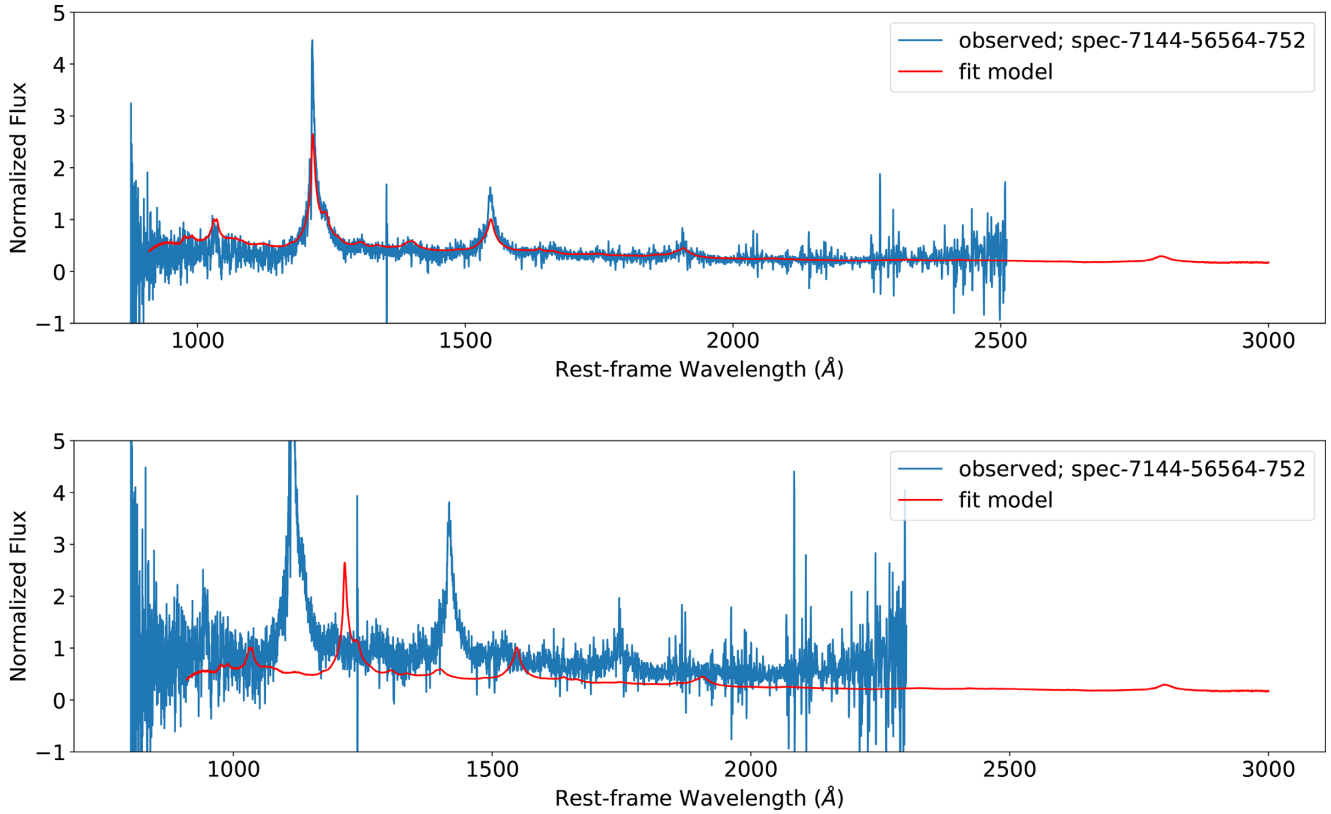

Figure 1. An example spectrum for our redshift estimation model. Red curve: the GP model mean for redshift estimation. Blue curve: the raw observed flux, after normalization at the range of 1216 ± 40 Å. (Top) At the maximum likelihood redshift of this quasar. (Bottom) At an incorrect quasar redshift. Note that the normalization of the quasar is incorrect. Normalizing in the 1216 ± 40 Å; region can introduce an additional penalty for incorrect redshifts.

continuum flux.[2] The variability of the line is encoded in the GP covariance function, see Fig. 3. We speculate that normalizing to Lyman α performs well because the strength of the Lyman α line minimizes the impact of instrumental noise in the normalizing region on the continuum normalization. As the Lyman α line is broad the normalization is also reasonably stable to small changes in $z_{\rm QSO}$.

We tried normalizing the quasar to the median continuum and to the CIV peak. Normalizing to the continuum led to complex unphysical structure in the learned covariance matrix and poor results. Normalizing to the CIV peak gave a tolerable covariance, but produced about a factor of 2 more redshift estimation failures than normalizing to the Lyman α peak.

During the testing phase, the observed flux x has to be normalized for each redshift possibility, as the region of observed spectrum which corresponds to the normalization region in the rest frame changes with assumed quasar redshift. We transform the spectrum as follows:

$$x \leftarrow x/\bar{x}(z_{\rm QSO})$$
$$\bar{x}(z_{\rm QSO}) = \text{median}\left[x(\text{rs}_{z_{\rm QSO}}(\lambda_{\rm OBS}) \in [1176\,\text{Å}, 1256\,\text{Å}])\right]. \quad (8)$$

This transformation is done separately for every redshift sample, $z_{\rm QSO}$. Thus, the normalization is redshift dependent and the likelihood depends only on the normalized flux. D^0 is again defined on the rest-frame wavelengths $\text{rs}_{z_{\rm QSO}}(\lambda)$ and the normalized flux $\tilde{y}$, which is the emission spectrum without any intervening DLAs.

[2]Interestingly, the automated quasar continuum estimator of Reiman et al. (2020) also normalizes continua using the height of the Lyman α peak

An incorrect normalization factor, $\bar{x}$, substantially changes the likelihood of the quasar. Thus, in most cases, the normalization factor is close to the true $\bar{x}(z_{\rm QSO_{true}})$ if and only if the $z_{\rm QSO} = z_{\rm QSOtrue}$, inducing an additional penalty in a $z_{\rm QSO}$ sample which is not close to the true quasar redshift. However, the roughly flat shape of the average quasar continuum means that fitting different emission peaks to Lyman α still produces a plausible normalization. Fig. 1 illustrates such an incorrect normalization from choosing a wrong $z_{\rm QSO}$.

3.5 Redshift estimation model summary

Combining all modelling decisions, the model prior for an observed QSO emission is

$$\begin{aligned} p(\tilde{y} = x \mid \text{D}^0, z_{\rm QSO}) &= \mathcal{N}\left(\frac{x}{\bar{x}(z_{\rm QSO})}; \mu \circ \text{rs}_{z_{\rm QSO}}, K \circ \text{rs}_{z_{\rm QSO}} + \frac{K_N}{\bar{x}(z_{\rm QSO})^2}\right) \\ &\times \prod_{\lambda \in \mathcal{X}_{\rm red}(z_{\rm QSO})} \mathcal{N}\left(\frac{x(\lambda_{OBS})}{\bar{x}(z_{\rm QSO})}; \mu_{\rm red}, \sigma^2_{\rm red} + \frac{\sigma^2_\lambda}{\bar{x}(z_{\rm QSO})^2}\right) \\ &\times \prod_{\lambda \in \mathcal{X}_{\rm blue}(z_{\rm QSO})} \mathcal{N}\left(\frac{x(\lambda_{OBS})}{\bar{x}(z_{\rm QSO})}; \mu_{\rm blue}, \sigma^2_{\rm blue} + \frac{\sigma^2_\lambda}{\bar{x}(z_{\rm QSO})^2}\right), \quad (9) \end{aligned}$$

where $\mathcal{X}_{\rm red}(z_{\rm QSO})$ are the set of observed wavelengths which fall outside of the Gaussian process model when transformed into a rest frame of $z_{\rm QSO}$. By sampling from the parameter prior $p(z_{\rm QSO})$, this model prior serves as a likelihood function for a QSO observation being at a given $z_{\rm QSO}$.

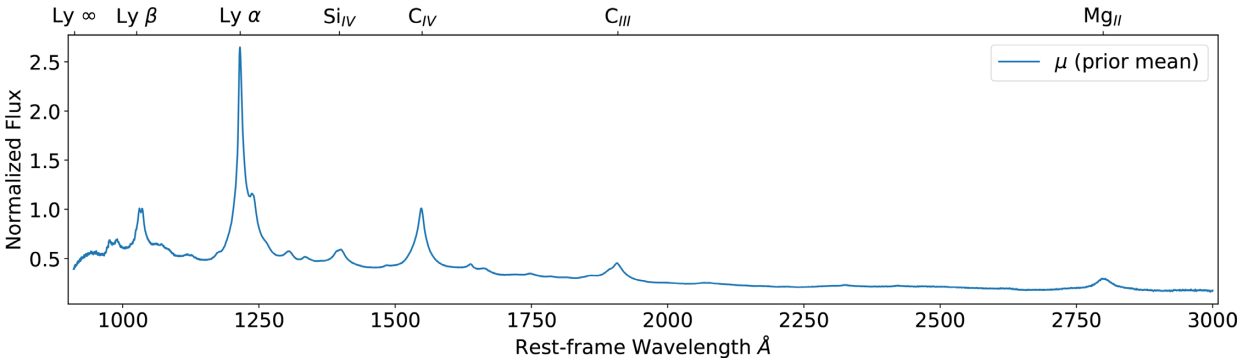

Figure 2. The estimated mean vector for our rest-frame quasar model, found by taking the mean value for each interpolated value across all rest-frame spectra in the training set. The rest-frame locations of common emission lines are shown in the upper axis.

The first $\mathcal{N}$ is the density of a Gaussian process, evaluated on the observations that fall within the Gaussian process model. The last two $\mathcal{N}$ are standard normal densities on the scalar values of the observations that fall outside the Gaussian process model. The observed instrumental noise is normalized by $\bar{x}(z_{\rm QSO})^2$, so that equation (8) shows the noise kernel K_N after normalization. $(\mu \circ \mathrm{rs}_{z_{\rm QSO}}, K \circ \mathrm{rs}_{z_{\rm QSO}})$ denotes the mean function and covariance kernel in the quasar rest frame. The mean function and covariance function are only modelled within the range based on equation (4). At the testing phase, we thus only evaluate the GP likelihood of $x(\lambda)$ inside the modelling window. We use the quasi-random Halton sequence to generate 10^4 samples of $z_{\rm QSO}$ from our prior for $p(z_{\rm QSO})$.

3.6 Learning the flux mean vector and covariance

In this section, we describe how we learn μ and K of our GP model D^0. Both are discretized. That is, we model μ as a piecewise constant function whose 'pieces' are of fixed widths. Thus, its parametrization is as a vector of the mean values over each piece. K is similarly discretized as a matrix.

Each observed spectrum is transformed to the rest frame and the values interpolated to the mid-points of the piecewise constant representation. Each element of the μ vector is estimated as the mean of all available[3] rest-frame flux values at the same wavelength. The learned mean from the data is shown in Fig. 2, and clearly shows the expected series of metal emission lines.

To acquire the kernel matrix K, we assume the same likelihood as Garnett et al. (2017) except (for now) excluding the absorption noise:

$$p(\boldsymbol{Y} \mid \mathrm{D}^0, z_{\rm QSO}) = \prod_{i=1}^{N_{\rm spec}} \mathcal{N}(\tilde{y}_i; \mu \circ \mathrm{rs}_{z_{\rm QSO}}, K \circ \mathrm{rs}_{z_{\rm QSO}} + K_N), \qquad (10)$$

where $\boldsymbol{Y}$ represents the matrix of all observed flux measurements in the training set, each transformed into the rest frame on a standard grid. The covariance matrix K is learned via the low-rank decomposition

$$\boldsymbol{K} = \boldsymbol{M}\boldsymbol{M}^\top. \qquad (11)$$

$\boldsymbol{K}$ is the kernel $(K \circ \mathrm{rs}_{z_{\rm QSO}})$, conditioned on the rest-frame wavelength pixels we defined before, and $\boldsymbol{M}$ is an $(N_{\rm pixels} \times k)$ matrix, with $N_{\rm pixels} = 8\,361$ and $k = 20$.

[3]Some observations are missing or have instrumental noise variance larger than 4^2 and are omitted.

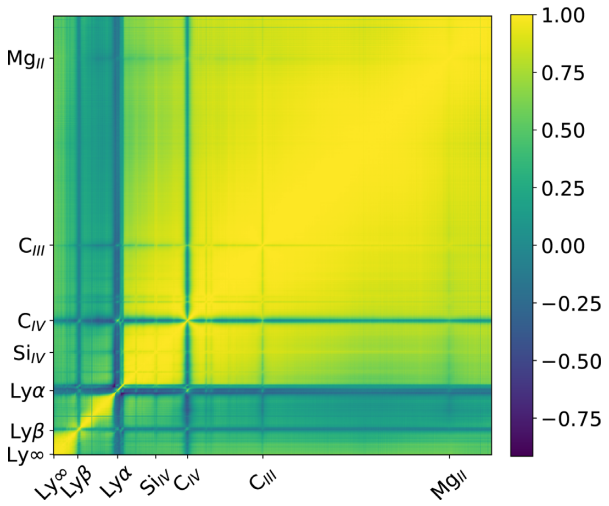

Figure 3. The trained correlation matrix $\boldsymbol{K}$, with the λ range from 910 to 3 000 Å. We have normalized the diagonal elements to be unity. The values in the matrix range from −1 to 1, representing the correlation between λ and λ′ in the QSO emission function.

Our kernel is trained by optimizing the values of $\boldsymbol{M}$ to maximize the likelihood given in equation (9). We use the first k principal components of $(\boldsymbol{Y} - \mu)$ as initial conditions. With the much larger model range (and thus matrices) trained in this paper, the MATLAB PCA function often failed to find principal components. This was due to substantial missing or noisy data at the red side of the training set. To allow the PCA to converge, we replaced all such data, represented in our data set by NaN, with the median value of the whole spectrum before taking the PCA. Although this kind of missing data imputation generally biases a PCA, in this case we are only using it as a starting point for our algorithm, and subsequently optimizing it away. Optimization is still done using the unmodified $\boldsymbol{Y}$ and uses the same unconstrained optimization as in our earlier papers, except without gradients of the absorption noise model.

Fig. 3 shows the learned kernel. The bottom left resembles the similar figure of Garnett et al. (2017), which was evaluated only in that range. The dark vertical lines in Fig. 3 show pixel areas which have strong correlations only in a narrow wavelength range. These areas drive the final redshift estimate and correspond to the locations of well-known major emission lines. Particularly strong are C IV (1549 Å), Lyman α (1216 Å), and O VI (1034 Å). Weaker signals are shown for Mg II (2799 Å), C III (1909 Å), Si IV (1397 Å), and C II

(1335 Å). Although Mg II is a famously reliable line (Hewett & Wild 2010), its presence in the correlation matrix is reduced because the emission line is low amplitude compared to the instrumental noise at long wavelengths. On the other hand, C IV, Lyman α and O VI are extremely strong emission lines and thus more visible. The width and variability in the Lyman α line position shows up as the width of the correlation band around 1216 Å. The similar width of the C IV line may be due to the code learning the correlation between C IV equivalent width and line blueshift (Gaskell 1982; Sulentic et al. 2007; Richards et al. 2011; Mason et al. 2017).

4 DLA FINDING MODEL

In this section, we describe how our quasar redshift estimator can be extended to find DLAs, while marginalizing out quasar redshift uncertainty. We take the model presented in Section 3 and combine it with the DLA model from Ho et al. (2020). The most important changes to the model are the inclusion of a model for Lyman-series absorbers along the line of sight to the quasar (Section 4.1) and an explicit model for DLAs (Section 4.2).

We do not use the uniform prior quasar redshift distribution from Section 3. Instead, we use as a prior a 150 bin histogram of $z_{\rm QSO}$ from the training data. We have checked explicitly that pure redshift estimation with this prior leads to similar results as the uniform prior, with some minor sampling artefacts at high redshift.

4.1 Lyman-series absorption

Following Ho et al. (2020), we supplement our instrumental noise model with an additional variance term to account for absorption from Lyman series lines, especially the Lyman α forest. We model Lyman series absorption as Gaussian noise with a redshift-dependent mean and variance, but no inter-pixel correlations. Our Gaussian process model for redshift estimation from Section 3 is thus modified by adding the diagonal absorption noise kernel K_A:

$$p(x|z_{\rm QSO}) = \mathcal{N}\left(x; \mu\circ{\rm rs}_{z_{\rm QSO}}, (K\circ{\rm rs}_{z_{\rm QSO}}) + K_A + K_N\right).$$

As Lyman α forest absorption is only possible in the region of the spectrum bluewards of the Lyman α line in the quasar rest frame, we include an indicator function in K_A, so that absorption is zero for $\lambda_{\rm REST} > 1216$ Å.

Evolution of the Lyman α forest flux with redshift is included by assuming the absorption noise has a power-law redshift dependence, so that K_A is given by

$$K_A(\lambda_{\rm REST}, \lambda'_{\rm REST}) = \delta(\lambda_{\rm REST} - \lambda'_{\rm REST}) \times I(\lambda_{\rm REST} < 1216)\left(1 - \exp\left(-\tau_0(1+z_{\rm Lya})^\beta\right) + c_0\right)^2,$$

where

$$z_{\rm Lya} = \frac{\lambda_{\rm REST}}{1216}\left(1+z_{\rm QSO}\right) - 1.$$

c_0, τ_0, and β are constants, and $z_{\rm Lya}$ is the redshift of Lyman α at the observed wavelength. Hence, our model depends on the redshift of the quasar as well as the redshift of Lyman α along the line of sight.

One unphysical feature of our absorption noise model is that, because Gaussian noise is symmetric, it assumes emission is as likely as absorption. This is particularly dangerous at high redshift, where the average absorption in a quasar spectrum is substantial. As we showed (Ho et al. 2020), we can account for this by modifying the quasar mean vector to match the observed mean flux of the Lyman α forest. We assume an effective optical depth $\tau_0(1+z_{\rm Lya})^\gamma$ following Kim et al. (2007):

$$a(z_{\rm Lya}) = \exp\left(-\tau_0(1+z_{\rm Lya})^\gamma\right) \tag{12}$$

$$= 0.0023 \times \exp\left(1+z_{\rm Lya}\right)^{3.65}, \tag{13}$$

We include absorption for the first six Lyman series lines, accounting for the different absorption coefficients. We account for the mean suppression from Lyman series absorption in our redshift-dependent noise model K_A. The complete GP model mean, written as a function of observed-frame wavelength $\lambda_{\rm obs}$, for each spectrum is thus

$$a(\lambda_{\rm obs}/\lambda_{\rm Ly\alpha} - 1) \times (\mu\circ{\rm rs}_{z_{\rm QSO}})(\lambda_{\rm obs}). \tag{14}$$

The parameters for the redshift-dependent component of the absorption noise vector were

$$c_0 = 0.3050; \quad \tau_0 = 1.6400\times10^{-4}; \quad \beta = 5.2714. \tag{15}$$

Once the absorption model is included, there are degeneracies between different hyperparameters of the GP kernel. This increases training time and means that the training does not technically converge. Instead it moves along a trough with the maximum likelihood changing by less than 0.1 per cent. Our trained model stopped training after 1500 minimization steps, although early iterations were trained to 3000 iterations with little difference in the kernel function.

4.2 DLA model

We introduce an alternate model for DLA spectra following Garnett et al. (2017). Either the DLA or no-DLA model is chosen by Bayesian model selection. The presence of a DLA is indicated by its Voigt profile, which includes absorption due to higher order Lyman lines:

$$\check{y}(\lambda_{\rm OBS}) = \tilde{y}(\lambda_{\rm OBS})\exp\left(-\tau({\rm rs}_{z_{\rm DLA}}(\lambda_{\rm OBS}); N_{\rm H\,I})\right).$$

Here, $\check{y}$ is the emission spectrum after DLA absorption and $\tau(\lambda; N_{\rm H\,I})$ is the Voigt profile for column density $N_{\rm H\,I}$ at wavelength λ. The DLA model ($\rm D^1$) has two parameters: the DLA redshift $z_{\rm DLA}$ and the DLA column density $N_{\rm H\,I}$. We take the prior redshift distribution of the DLA, $p(z_{\rm DLA}|{\rm D}^1, z_{\rm QSO})$, to be uniform between a region 3000 km s^{-1} redwards of the Lyman limit at 910 Å and 3000 km s^{-1} bluewards of $z_{\rm QSO}$.

The prior distribution over the column density, $p(N_{\rm H\,I}|{\rm D}^1)$, is modelled as a lognormal distribution. We use a kernel density estimate from the DR 9 sample, mixed with a uniform distribution (equation 51 of Garnett et al. 2017). We do not include the sub-DLA model of Ho et al. (2020).

4.3 Model inference

Our full model is

$$\begin{aligned} p(x, z_{\rm QSO}, {\rm D}^0) &= p(z_{\rm QSO}) \times \Pr({\rm D}^0 \mid z_{\rm QSO}) \\ &\quad\times,\ p(x \mid {\rm D}^0, z_{\rm QSO}) \\ p(x, z_{\rm QSO}, {\rm D}^1, z_{\rm DLA}, N_{\rm H\,I}) &= p(z_{\rm QSO}) \times \Pr({\rm D}^1 \mid z_{\rm QSO}) \\ &\quad\times p(z_{\rm DLA} \mid z_{\rm QSO}, {\rm D}^1) \times p(N_{\rm H\,I} \mid {\rm D}^1) \\ &\quad\times p(x \mid {\rm D}^1, z_{\rm QSO}, z_{\rm DLA}, N_{\rm H\,I}). \end{aligned}$$

$z_{\rm DLA}$ and $N_{\rm H\,I}$ can be marginalized out to obtain

$$p(x, z_{\rm QSO}, {\rm D}^1) = \int\int p(x, z_{\rm QSO}, {\rm D}^1, z_{\rm DLA}, N_{\rm H\,I})\, dz_{\rm DLA}\, dN_{\rm H\,I}.$$

We are particularly interested in $\Pr({\rm D}^1 \mid x)$, the probability of a DLA given the observed spectrum, and $p(z_{\rm QSO}|x)$, the distribution

of the quasar redshift given the observed spectrum. We calculate these conditional marginal distributions as follows.

$$p(\mathrm{D}^1, x) = \int p(x, z_{\mathrm{QSO}}, \mathrm{D}^1)\, \mathrm{d}z_{\mathrm{QSO}}, \tag{16}$$

$$p(\mathrm{D}^0, x) = \int p(x, z_{\mathrm{QSO}}, \mathrm{D}^0)\, \mathrm{d}z_{\mathrm{QSO}}, \tag{17}$$

$$\Pr(\mathrm{D}^1 \mid x) = \frac{p(\mathrm{D}^1, x)}{p(\mathrm{D}^1, x) + p(\mathrm{D}^0, x)}, \tag{18}$$

and

$$p(z_{\mathrm{QSO}} \mid x) \propto p(x, z_{\mathrm{QSO}}, \mathrm{D}^1) + p(x, z_{\mathrm{QSO}}, \mathrm{D}^0), \tag{19}$$

where the constant of proportionality in the last line makes $p(z_{\mathrm{QSO}}|x)$ integrate to 1 over z_{QSO}.

Estimating the probability of a DLA requires a three-dimensional integral over $\{z_{\mathrm{QSO}}, z_{\mathrm{DLA}}, N_{\mathrm{H\,I}}\}$ for $p(\mathrm{D}^1, x)$ and a one-dimensional integral over $\{z_{\mathrm{QSO}}\}$ for $p(\mathrm{D}^0, x)$. As in Section 3, we use the quasi-random Halton sequence to generate one- or three-dimensional points as samples over the unit cube. However, reflecting the higher dimensionality of our parameter space we draw 10^5 samples per quasar instead of 10^4. We then transform them by the relevant inverse cumulatives to generate samples from $p(z_{\mathrm{QSO}})$ or $p(z_{\mathrm{QSO}}, z_{\mathrm{DLA}}, N_{\mathrm{H\,I}})$ from which the integrals can be numerically approximated (as the integrals can be transformed into expectations with respect to these sampling distributions). In this way, the likelihood of a DLA can be estimated without knowledge of z_{QSO}.

4.4 Model parametrization and priors

The full model requires the specifications of the following components. In the quasar rest frame:

(i) μ: the mean quasar emission spectrum and
(ii) K: the kernel of the Gaussian process for the emission.

In the redshifted observer frame:

(i) K_A: the diagonal non-DLA absorption variance and
(ii) K_N: the diagonal instrument noise variance.

Priors are given for

(i) $p(z_{\mathrm{QSO}})$: the redshift of a quasar,
(ii) $p(N_{\mathrm{H\,I}}|\mathrm{D}^1)$: the column density of the DLA, and
(iii) $p(z_{\mathrm{DLA}}|\mathrm{D}^1, z_{\mathrm{QSO}})$: the DLA redshift distribution.

5 TRAINING AND VALIDATION DATA

The training set to learn our GP model D^0 for z_{QSO} estimate consists of the spectra observed by SDSS DR9. For DLA finding we also removed DLAs labelled in Lee et al. (2013). The validation data consisted of SDSS DR12, comprising 297 301 quasar spectra. The following spectra were removed from both the training and validation set:

(i) $z_{\mathrm{VI}} < 2.15$: quasars with redshifts lower than 2.15.
(ii) BAL: quasars where SDSS found BLAs.
(iii) Spectra with less than 400 detected pixels.
(iv) ZWARNING: spectra whose analysis by the SDSS pipeline flagged warnings. These spectra are usually not quasars, but represent some instrumental problem. We kept extremely noisy spectra with the TOO_MANY_OUTLIERS flag.

After these cuts, the remaining sightline catalogue is 158 979 quasars. Given that the purpose of this paper is redshift estimation, it may seem circular to filter quasars with $z < 2.15$ from testing. However, these quasars do not contain DLAs, nor are they useful for Lyman α BAO. We examine these spectra further in Section 6.1.1 and show that our trained model still works reasonably well as long as the Lyman α emission peak (which we use for normalization) is inside the observed band, that is for $z \geq 1.9$.

Training the model requires a redshift estimate for the training data. Here, we use the SDSS visual inspection redshift as it is available for the highest quasar fraction in the sample (Pâris et al. 2018). Note, however, that visual inspection redshifts are not required. The model merely requires some redshift estimate. Future iterations could be trained using, for example, the DR12 redshift outputs of this paper, making the model fully self-hosting.

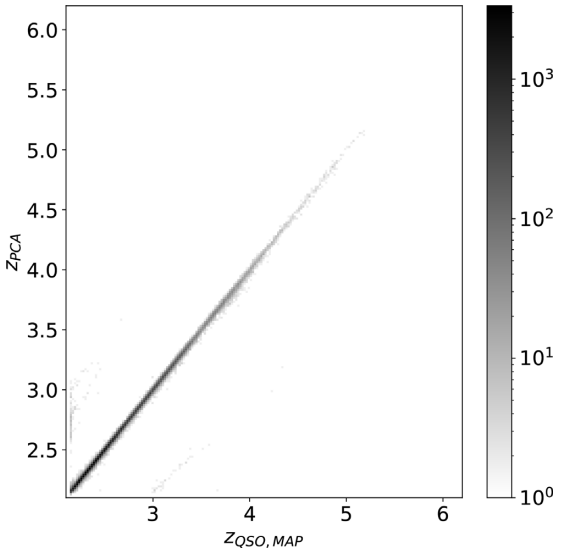

Figure 4. The MAP prediction of our catalogue, z_{QSO}, versus the PCA redshift z_{PCA} from the SDSS catalogue. The grey-scale bar shows the number of quasars in each bin, using a logarithmic scale. The diagonal line in the middle of the plot shows a correct redshift estimation. Other diagonal lines correspond to occasional line fitting mistakes of our code.

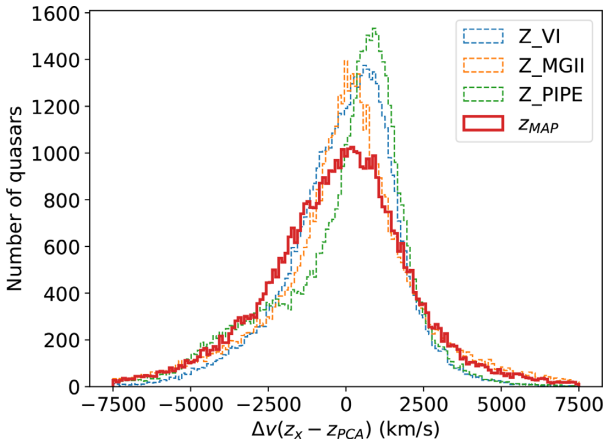

Figure 5. A histogram showing differences between the MAP prediction of z_{QSO} from our catalogue (z_{MAP}) and different z_{QSO} estimation techniques present in the SDSS catalogue. All methods are compared to the PCA based redshift, z_{PCA}. We show results for only the 49 776 quasars with redshift estimates from all SDSS methods.

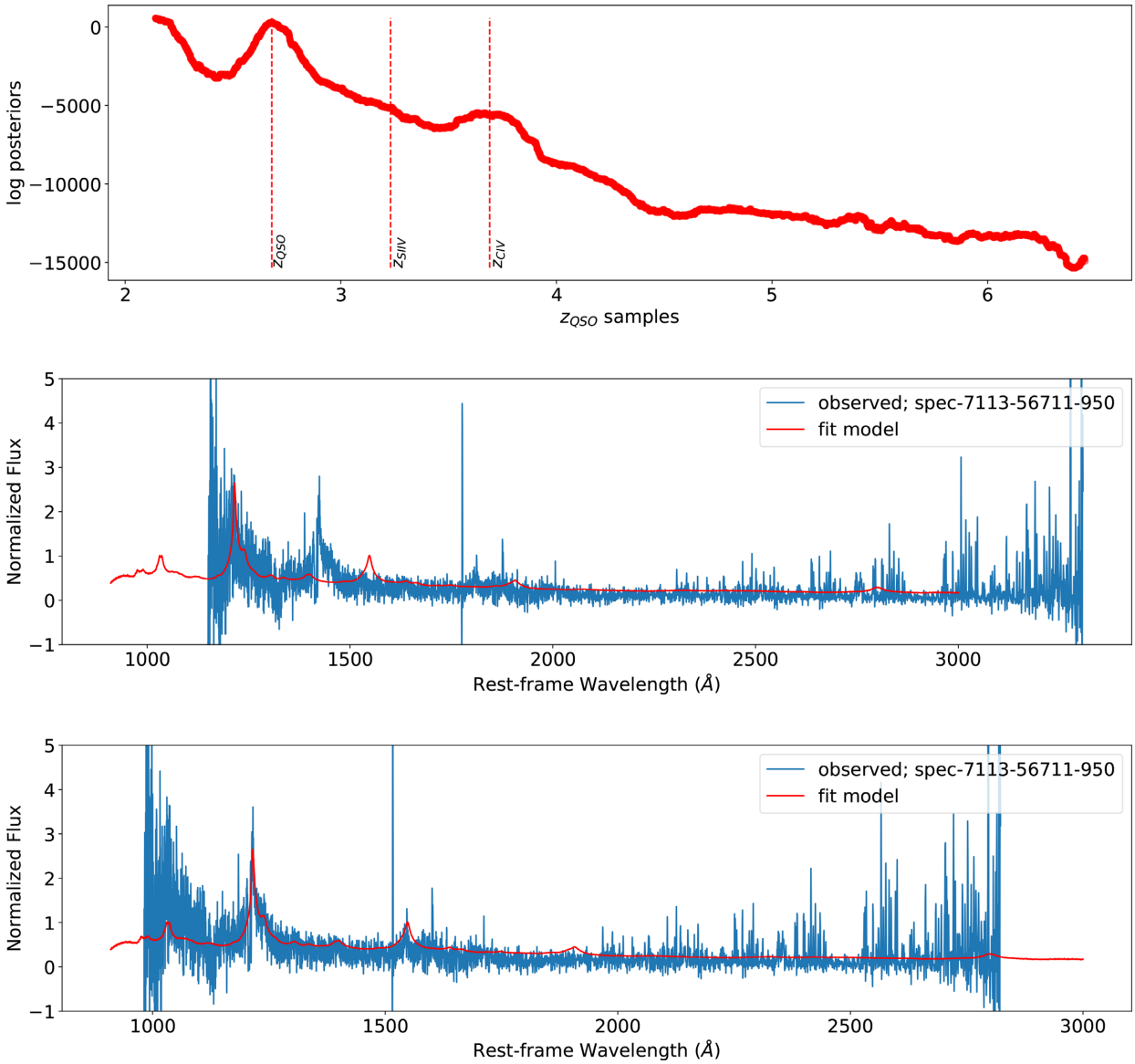

Figure 6. (Top) The sample posterior $p(z_{\rm QSO}|x, {\rm D}^0)$ for a QSO with thingID = 544031279. The catalogue redshift is labelled as $z_{\rm QSO}$. Vertical dashed lines indicate the redshifts associated with samples at particular emission peaks. For example, the redshift resulting from trying to fit the true Lyman α peak on to the observed C IV peak is shown as $z_{\rm C\,IV}$. (Middle) The rest-frame spectrum using $z_{\rm MAP}$. (Bottom) The rest-frame spectrum using the SDSS visual inspection redshift $z_{\rm VI}$. We use $z_{\rm VI}$ as it is the method with the lowest failure rate. The MAP value of our catalogue fits the Lyman α peak with what is really O VI.

6 RESULTS

In this section, we describe the results of our algorithm run on the SDSS DR12Q data set. Section 6.1 describes the results when estimating only quasar redshift. Section 6.2 also describes the results of our DLA finding.

6.1 Redshift estimation

In this section, we apply our QSO redshift model $\rm D^0$ to SDSS DR12. We validate our ability to predict quasar redshift, $z_{\rm QSO}$. Although our model is fully Bayesian, we need a point estimate to compare to the SDSS catalogue redshift. We use the *maximum a posteriori* (MAP) of the sample posterior $p(z_{\rm QSO}|x, {\rm D}^0)$, which is equivalent to the maximum likelihood estimate (MLE) because we use a uniform prior for $p(z_{\rm QSO})$. We thus report the $z_{\rm QSO}$ sample with the highest likelihood

$$z_{\rm MAP} = \arg\max_{z_{{\rm QSO}i}} \; p(\tilde{y}(z_{{\rm QSO}i}) \mid {\rm D}^0, z_{{\rm QSO}i}), \tag{20}$$

where $z_{{\rm QSO}i}$ is the ith Halton sequence sample. The instrumental noise variance depends on $z_{{\rm QSO}i}$ via normalization.

In Fig. 4, we compare the MAP estimate of our catalogue, $z_{\rm MAP}$, to the reported PCA redshift $z_{\rm PCA}$ in SDSS DR12. The two are generally in good agreement, as shown by the large number of quasars on the plot diagonal. There are a small number of cases where our model fits Lyman α using another emission peak, visible as the secondary lines above and below the main diagonal (note that Fig. 4 uses a logarithmic scale). The above-diagonal line corresponds to Lyman α peaks being fit by O VI emission. This line is broad because O VI is in the Lyman α forest and so has large variance in our model. The below-diagonal line, which is narrower, corresponds to Lyman α peaks fit with C IV emission. There are also a few objects, of a

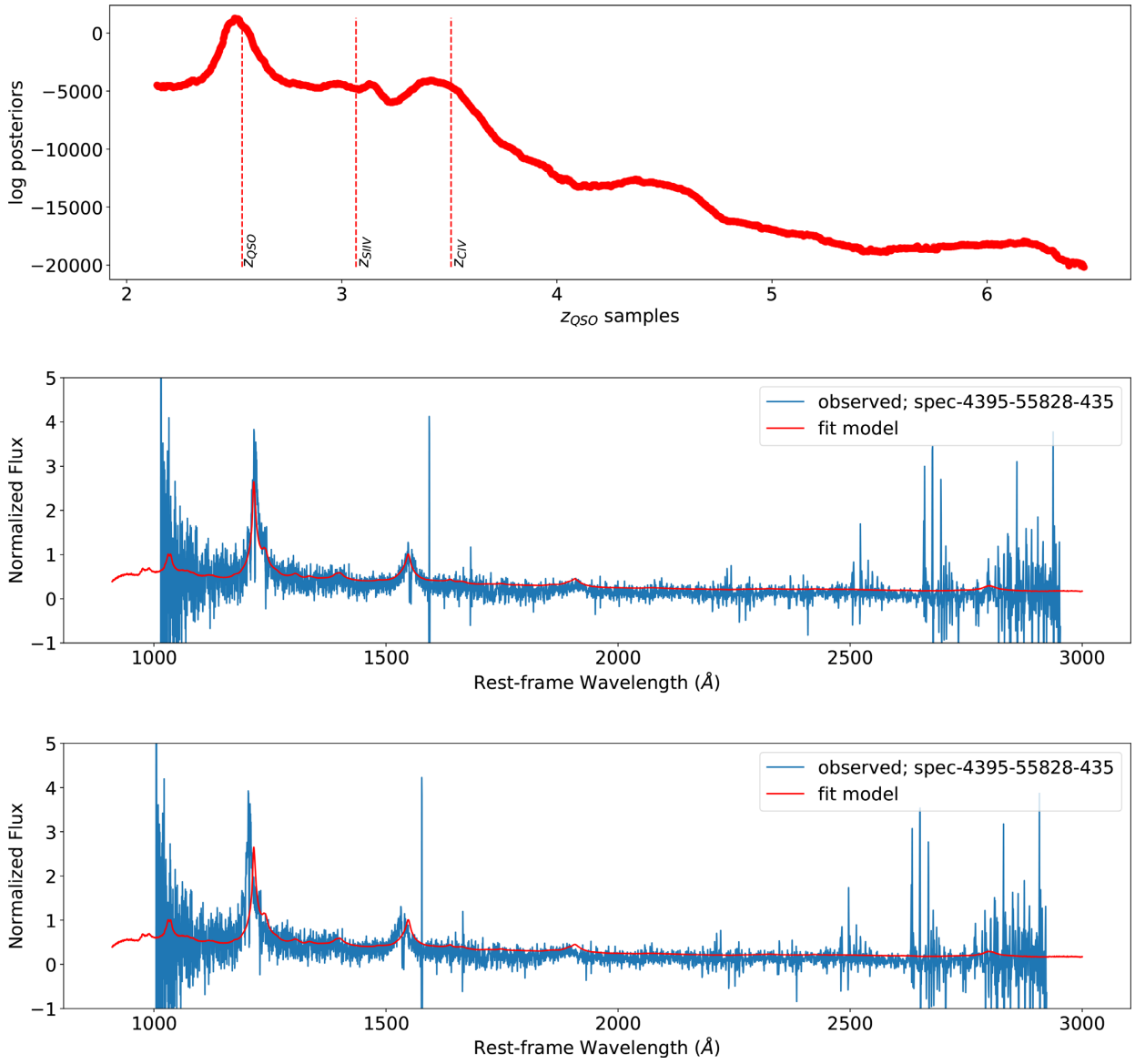

Figure 7. (Top) The sample posterior $p(z_{\rm QSO}|x, {\rm D}^0)$ for a QSO with thingID = 27885089. The catalogue redshift is labelled as $z_{\rm QSO}$. Vertical dashed lines indicate the redshifts associated with samples at particular emission peaks. For example, the redshift resulting from trying to fit the true Lyman α peak on to the observed C IV peak is shown as $z_{\rm C\,IV}$. (Middle) The rest-frame spectrum using $z_{\rm MAP}$. (Bottom) The rest-frame spectrum using the SDSS visual inspection redshift $z_{\rm VI}$. We use $z_{\rm VI}$ as it is the method with the lowest failure rate. $z_{\rm MAP}$ appears to produce a better fit than $z_{\rm VI}$.

density too low to be visible on the plot, where the code fits the O VI emission line to C IV. The rate at which our redshift estimation fails is low. Comparing to the PCA redshift we find that $|z_{\rm PCA} - z_{\rm MAP}| > 0.5$ for 0.38 per cent, which is 603 out of 158 560 quasar spectra. Comparing to the visual inspection redshift z_{VI} gives similar results: $|z_{VI} - z_{\rm MAP}| > 0.5$ for 645 of 158 979 spectra. For the more stringent bound of $|z_{\rm VI} - z_{\rm MAP}| > 0.05$, the misfit rate rises to 0.99 per cent. Other redshift measurements performed similarly, with $z_{\rm C\,IV}$ having the lowest misfit rate (0.35 per cent) and $z_{\rm PIPE}$ the highest (0.43 per cent).

Fig. 5 compares to other redshift estimation methods used in SDSS, following Fig. 7 of Pâris et al. (2017). We show results only for the 49 776 quasars with redshift estimates from all SDSS methods. Overall our technique performs similarly to the others. It is complementary in that it prefers lower redshifts than the PCA model $z_{\rm PCA}$, while other methods prefer a generally higher redshift.

Our method has a median difference in redshift with $z_{\rm PCA}$ of -117 km s^{-1}. The equivalent median differences between $z_{\rm PCA}$ and other methods are $z_{\rm VI}$: 128 km s^{-1}, $z_{\rm Mg\,II}$: 73 km s^{-1}, $z_{\rm PIPE}$: 380 km s^{-1}. Our technique is thus competitive in this metric. The standard deviation of this dispersion with $z_{\rm PCA}$ is 17 000 km s^{-1}. The other methods score substantially better: $z_{\rm VI}$: 1800 km s^{-1}, $z_{\rm Mg\,II}$: 2500 km s^{-1}, $z_{\rm PIPE}$: 12 000 km s^{-1}. For both our method and $z_{\rm PIPE}$, the large standard deviations are driven by the relatively large fraction of outliers, i.e., catastrophic failures of redshift determination. The inter-quartile range for each method shows a measurement of dispersion which is not affected by these failures. We have $z_{\rm MAP}$: 3000 km s^{-1}, $z_{\rm VI}$: 1200 km s^{-1}, $z_{\rm Mg\,II}$: 1700 km s^{-1}, $z_{\rm PIPE}$: 1600 km s^{-1}. Our redshift estimation method thus produces a larger dispersion than the other methods.

We have visually inspected a subsample of the spectra where our catalogue has a dramatically incorrect redshift. Fig. 6 shows

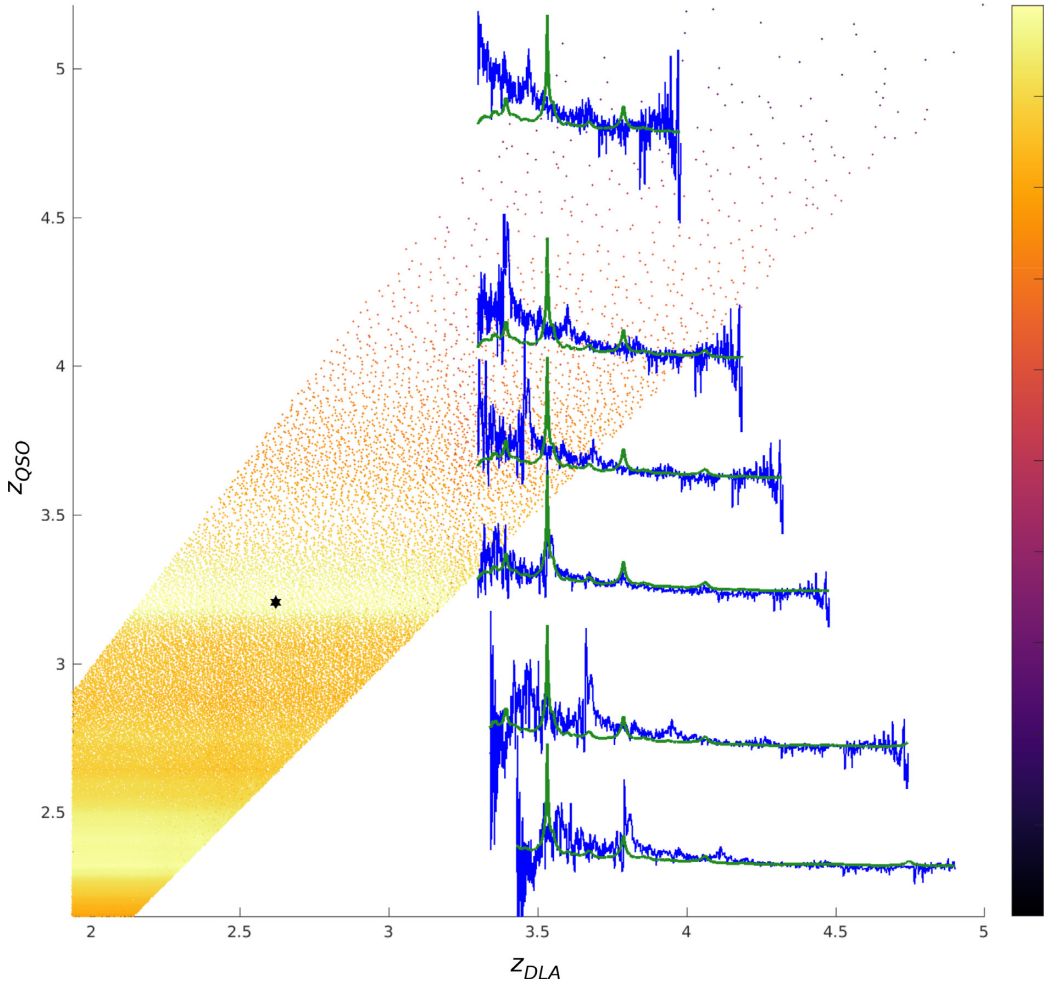

Figure 8. Examples of Halton sequence sampling for $z_{\rm QSO}$, $z_{\rm DLA}$, and $N_{\rm H\,I}$. Samples across parameter space Θ project out $N_{\rm H\,I}$ on to the $z_{\rm QSO}$ and $z_{\rm DLA}$ plane. The best sample (at $z_{\rm QSO} = 2.309$) is shown by a black star. Colours estimate the posterior log-likelihood of D^1 for each point. $z_{\rm DLA}$ is drawn uniformly while $z_{\rm QSO}$ is taken from an empirical distribution. This particular quasar has a bimodal likelihood for $z_{\rm QSO}$, where the second, lower, peak corresponds to the code fitting the Lyman α peak at O VI. Though estimates of $z_{\rm DLA}$ are drawn uniformly, the DLA cannot appear redwards of the quasar or bluewards of the Lyman α peak, and so are not sampled from these regions. Shown for reference in green are illustrations of the given quasar and rest-frame mean prediction in the rest frame for $z_{\rm QSO}$ sampled at 2.3, 2.7, 3.22, 3.6, 4.0, 4.75. Normalizations for the spectra are 1.93, 1.43, 1.80, 1.06, 0.78, 0.51, respectively.

one such example. Here, the likelihood peaks at very low redshift, because the code believes that a noise peak near the O VI emission line is the Lyman α peak, and this overwhelms the otherwise poor fit to the spectrum. Note that there is a peak in the likelihood at the correct redshift, with almost the same probability, so a full Bayesian analysis would be closer to the true value. This spectrum, like most of those where the code confuses O VI for Lyman α, shows unusually noisy data with an oscillatory feature which exceeds the expected pipeline noise at the far blue end of the observed data, possibly related to the data reduction systematic identified by Lan et al. (2018). Spectra where the code confuses C IV for Lyman α often have unusually weak Lyman α peaks relative to their C IV emission.

There are also spectra in our catalogue where our method produces what looks visually like a better fit to the observed spectrum than $z_{\rm VI}$. Fig. 7 shows an example, where the catalogue $z_{\rm VI}$ redshift does not quite match the location of the C IV peak, possibly as an attempt to fit to noise near the Mg II emission line. Our method estimates redshift as $z_{\rm QSO} = 2.501$. Redshift estimates from the SDSS catalogue are $z_{\rm VI} = 2.538$, $z_{\rm PIPE} = 2.507$, $z_{\rm PCA} = 2.511$. $z_{\rm Mg\,II}$ was not available. In this case, $z_{\rm VI}$ is an outlier, and our model is in reasonable agreement with $z_{\rm PIPE}$. We note that the position of the C IV emission peak shown in the figure is from the mean model, and thus automatically includes the average C IV blueshift from the rest-frame emission (Hewett & Wild 2010; Richards et al. 2011).

6.1.1 *Validating the model at lower redshifts*

In this section, we validate the behaviour of our GP model D^0 on quasars with redshift outside the redshift range containing DLAs. We place a uniform prior on $z_{\rm QSO}$ as in equation (3), but we modify the lower bound to be $z_{\rm QSO\,min} = 1.9$. We select the test set as described in Section 5 except that we modify the range of $z_{\rm QSO}$ to be $1.9 < z_{\rm QSO} < 2.15$. The new sample size is 16 013 quasars. We do not retrain the model.

The catastrophic misfit rate for $|z_{\rm VI} - z_{\rm MAP}| > 0.5$ is 3.3 per cent. The error, as expected, is much larger than the results for spectra with $2.15 \leq z_{\rm QSO}$, as the Lyman α peak is now located at a lower observed-frame wavelength, where instrumental noise is larger. Since we normalize by the height of the Lyman α peak, noise in this region can easily lead us to produce an inaccurate continuum.

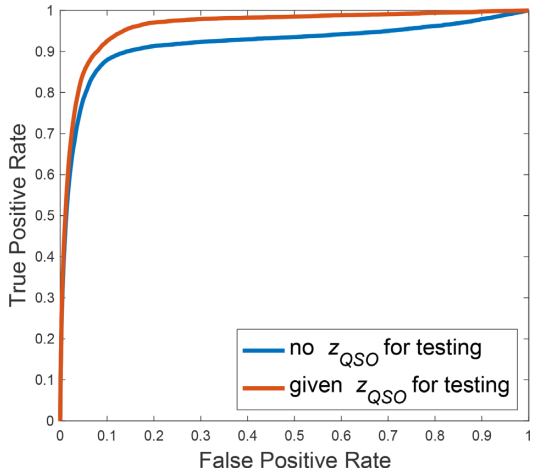

Figure 9. ROC curve for DLA estimation from our catalogue estimating $z_{\rm QSO}$ (blue), and from the catalogue of Ho et al. (2020) with $z_{\rm QSO}$ given (red). The AUC with full integration is 0.9192. The AUC from Ho et al. (2020) is 0.9624. The ROC is taken over all 158 821 applicable quasars in the DR12 data set. Ground truth is the best 2/3 catalogue for DR12, described in the text.

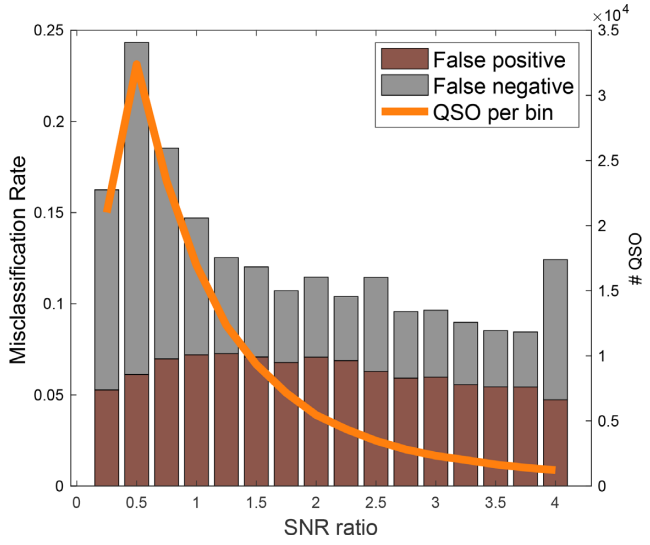

Figure 10. Error rate plotted as a function of signal-to-noise ratio. Curve and right y-axis shows the total number of quasars in each signal-to-noise bin, while the left y-axis shows the error rate. We consider that a spectrum has a DLA in our catalogue if $p(DLA) > 0.9$. Low SNR spectra have a higher level of false negatives. Ground truth is the best 2/3 catalogue for DR12, described in the text.

This normalization also leads to a natural minimum quasar redshift possible with our method at $z_{\rm QSOmin} = 1.9$, below which the Lyman α peak has not yet redshifted into the observation window of BOSS optical spectra (3650−10 400 Å). We can achieve slightly improved results for lower $z_{\rm QSO}$ samples by using a GP model trained by normalizing on C IV peak, 1549 ± 40 Å;. Here, the misfit rate was 2.8 per cent for $|z_{\rm VI} - z_{\rm MAP}| > 0.5$. However, normalizing to C IV performs substantially less well for quasars with $z_{\rm QSO} > 2.15$.

6.2 DLA finding

We now show our DLA catalogue computed with a marginalized $z_{\rm QSO}$. We have checked explicitly that redshift estimation is similar in this catalogue to the pure redshift estimation model discussed in Section 6.1. A two-dimensional projection showing $z_{\rm QSO}$ and $z_{\rm DLA}$, for an example quasar with a DLA, can be seen in Fig. 8. The mean over the product of each Bayes factor with each model prior for different $z_{\rm QSO}$ yields our posterior odds, which can be normalized to give our desired model posteriors $\Pr(\mathrm{D}^1|\mathcal{D})$ and $\Pr(\mathrm{D}^0|\mathcal{D})$.

6.2.1 *Best 2/3 DLA catalogue*

To compare our results to a single 'ground truth' DLA catalogue, we follow a procedure similar to that used to generate the DR9 concordance DLA catalogue (Lee et al. 2013). Aside from our work, there are three extant DR12 catalogues. These are Parks et al. (2018)[4] (based on a neural network), a DR12 catalogue generated using the template matching method of Noterdaeme et al. (2012) and the DR12 visual survey (Pâris et al. 2017).[5] Each method produces a slightly different DLA catalogue, differing by up to ∼10 per cent. However, by taking only DLAs which occur in 2/3 catalogues, we hope to produce a relatively pure sample.

To demonstrate our model effectiveness, we order each spectrum by its log posterior odds of D^1, with associated DLA information. Spectra which are assigned a DLA by our best 2/3 catalogue should appear at the top of this ordering as most probable. Fig. 9 shows the receiver-operating characteristic (ROC) plot of each method, comparing our current method integrating over $z_{\rm QSO}$ to a model with $z_{\rm QSO}$ assumed known (Ho et al. 2020). The AUC between our $z_{\rm QSO}$ marginalizing catalogue with full $z_{\rm QSO}$ integration and the best 2/3 is 0.9192. The AUC with known redshifts is 0.9624. The AUC between our current catalogue and that with known redshifts was 0.914, similar to the AUC between the $z_{\rm QSO}$ catalogue and the best 2/3.

Our method performs moderately less well than a similar integration task where $z_{\rm QSO}$ is given. This is not surprising, as the integration task without $z_{\rm QSO}$ is more difficult. While both models ultimately recover similar information, the full integration method estimates DLAs with less certainty, leading to a true positive rate which is worse by a few per cent. When a DLA is correctly identified the MAP DLA redshift and column density is similar to our previous papers, exhibiting no noticeable preference for higher or lower column densities. In particular, there are several instances where the DLA redshift is correctly determined despite the quasar redshift being incorrect.[6]

If our lower true positive rate is due simply to the increased difficulty of the problem, the presence of spectral noise should reduce the ability of our model to determine $z_{\rm QSO}$. Fig. 10 shows the error rate as a function of our catalogue's signal-to-noise ratio. Signal to noise was taken over as much of each quasar as could possibly sit in the rest frame, as a per-pixel mean of the flux over the square root of the noise variance. Also shown is the overall frequency of quasars per bin. Our false negative rate is indeed higher by a factor of 2 at low SNR. This may indicate that false negatives occur because there is not enough information for the model to make a solid detection. It is also possible that that these are not, in fact, real DLAs, and the low

[4]We include sub-DLAs from this catalogue so that the minimum column density from all catalogues is 10^{20} cm^{-2}, as the other catalogues.
[5]All DLAs of which we assign an arbitrary column density of 2×10^{20} cm^{-2}.
[6]This is possible because the transformation between observed frame and DLA frame does not depend on the quasar rest frame, as long as the measured $z_{\rm QSO}$ allows for a DLA in the observed region.

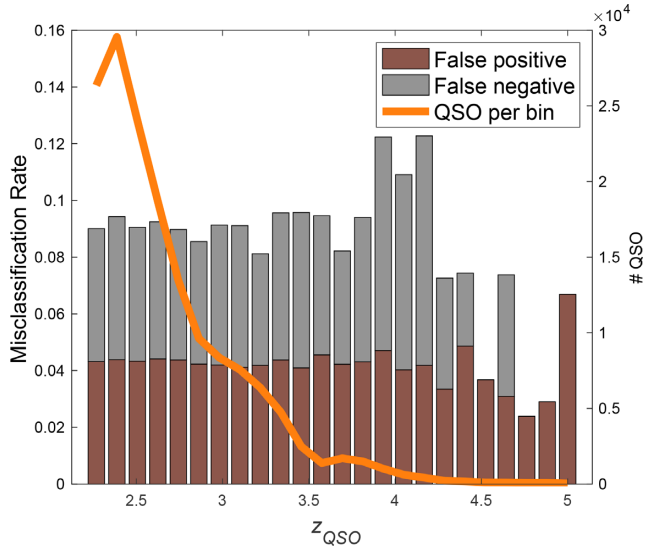

Figure 11. Error rate plotted as a function of quasar redshift. Curve and right y-axis shows the total number of quasars in each redshift bin, while the left y-axis shows the error rate. We consider that a spectrum has a DLA in our catalogue if $p(DLA) > 0.9$. Ground truth is the best 2/3 catalogue for DR12, described in the text.

signal-to-noise ratio was causing a slightly incorrect pipeline z_{QSO} which was misleading our previous DLA algorithm.

We have visually inspected a sample of low signal-to-noise spectra with false positive DLAs and poor redshift estimation. There are several examples where only 0–1 emission peaks emerge from the noise. Our false positives commonly occur in spectra where, if one takes the SDSS pipeline redshift as ground truth, one observes a Lyman break with noise at 700−800 Å. Our pipeline instead fits the O VI emission peak with Lyman α and interprets the break as a DLA. We suspect that most of these cases are indeed false positives, but obtaining reliable results from SNR < 1 will always be challenging.

Fig. 11 shows the error rate as a function of quasar redshift. The false positive rate is roughly independent of redshift, while the false negative rate is constant until $z = 3.6$. At $z > 4.2$ the false negative rate approaches zero. However, there are very few DLAs detected at this redshift in the best 2/3 catalogue. For $z = 3.7$–4.0 the false negative rate increases noticeably. In this redshift range, the Lyman break at 910 Å redshifts into the observed SDSS band, and it may be that our redshift estimation was confused by the presence of this feature in the spectrum.

7 CONCLUSION

We have extended our Gaussian process based code for finding DLAs in SDSS quasars to situations where the quasar redshift is not known. This required extending the Gaussian process range to encompass more emission lines and thus get a more reliable z_{QSO} estimate. It was also necessary to augment the model to include a likelihood component for all observations, even those which are outside the range of the Gaussian process, so that the probabilities are comparable for the same spectrum across multiple redshifts.

We first estimated the redshift of the SDSS DR12 sample, showing that our redshift labelling is competitive to existing redshift estimation. Large redshift misestimation was reasonably rare. Our redshift estimate differs from the PCA redshift by >0.5 for 603 quasars out of $\sim 1.6 \times 10^5$. The median redshift error of our method compared to other SDSS redshift estimates was ~ 100 km s^{-1}. We used our improved model to find DLAs while marginalizing over uncertainties in the quasar redshift. We detected a few per cent fewer DLAs at high confidence than our earlier methods (AUC drops from 0.96 to 0.91) , especially in noisy spectra where estimation is more difficult.

The computation time for the pure redshift estimation model is ~ 1.5 s per spectrum on a 48-core AWS EC2 machine, while finding DLAs takes ~ 60 s per quasar.

There are a few ways in which the redshift estimation present here may be improved. Our choice of normalization (the Lyman α peak) makes low-redshift quasars hard to classify correctly. In future work it might be better to incorporate normalization directly into the Bayesian model as an extra parameter. We may also have reached the limits of the Halton sequence based quasi Monte Carlo integrator we have used since Garnett et al. (2017). Future work may find it necessary to switch to a more targeted integrator based on variational or Markov chain Monte Carlo methods.

ACKNOWLEDGEMENTS

We thank Yongda Zhu and Marie Wingyee Lau for useful conversations. SB was supported by NSF grant AST-1817256. RG was supported by the NSF under award numbers IIS-1939677, OAC-1940224, and IIS-1845434. SB and RG were supported by an Amazon.com Machine Learning Research Award, which also provided computing time. CRS was supported in part by NSF grant (IIS-1510741). Computing time was also provided by UCR HPCC.

DATA AVAILABILITY

All the code to reproduce the data products is available in our GITHUB repo: https://github.com/sbird/gp_qso_redshift. The final data products are available in this Google Drive: http://tiny.cc/gp_zestimation_catalogue, including a MAT (HDF5) catalogue and a JSON catalogue.

REFERENCES

Alam S. et al., 2015, ApJS, 219, 12
Bird S., Vogelsberger M., Haehnelt M., Sijacki D., Genel S., Torrey P., Springel V., Hernquist L., 2014, MNRAS, 445, 2313
Bird S., Garnett R., Ho S., 2017, MNRAS, 466, 2111
Dawson K. S. et al., 2013, AJ, 145, 10
Dawson K. S. et al., 2016, AJ, 151, 44
DESI Collaboration, 2016, preprint (arXiv:1611.00036)
Eisenstein D. J. et al., 2011, AJ, 142, 72
Garnett R., Ho S., Bird S., Schneider J., 2017, MNRAS, 472, 1850
Gaskell C. M., 1982, ApJ, 263, 79
Haehnelt M. G., Steinmetz M., Rauch M., 1998, ApJ, 495, 647
Hewett P. C., Wild V., 2010, MNRAS, 405, 2302
Ho M.-F., Bird S., Garnett R., 2020, MNRAS, 496, 5436
Kim T.-S., Bolton J. S., Viel M., Haehnelt M. G., Carswell R. F., 2007, MNRAS, 382, 1657
Lan T.-W., Ménard B., Baron D., Johnson S., Poznanski D., Prochaska J. X., O'Meara J. M., 2018, MNRAS, 477, 3520
Lee K.-G. et al., 2013, AJ, 145, 69
Mason M., Brotherton M. S., Myers A., 2017, MNRAS, 469, 4675
Noterdaeme P. et al., 2012, A&A, 547, L1
Pâris I. et al., 2017, A&A, 597, A79
Pâris I. et al., 2018, A&A, 613, A51
Parks D., Prochaska J. X., Dong S., Cai Z., 2018, MNRAS, 476, 1151
Prochaska J. X., Wolfe A. M., 1997, ApJ, 487, 73

Reiman D. M., Tamanas J., Prochaska J. X., Ďurovčíková D., 2020, preprint (arXiv:2006.00615)
Richards G. T. et al., 2011, AJ, 141, 167
Schneider D. P. et al., 2010, AJ, 139, 2360
Shen Y. et al., 2016, ApJ, 831, 7
Sulentic J. W., Bachev R., Marziani P., Negrete C. A., Dultzin D., 2007, ApJ, 666, 757
Williams C. K., Rasmussen C. E., 2006, Gaussian Processes for Machine Learning, Vol. 2. MIT Press, Cambridge, MA
Wolfe A. M., Turnshek D. A., Smith H. E., Cohen R. D., 1986, ApJS, 61, 249

This paper has been typeset from a TEX/LATEX file prepared by the author.