Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **472,** 1850–1865 (2017) doi:10.1093/mnras/stx1958
Advance Access publication 2017 August 2

Detecting damped Ly α absorbers with Gaussian processes

Roman Garnett,[1]★ Shirley Ho,[2] Simeon Bird[3] and Jeff Schneider[4]

[1]*Department of Computer Science and Engineering, Washington University in St Louis, One Brookings Drive, St Louis, MO 63130, USA*
[2]*Department of Physics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA*
[3]*Department of Physics & Astronomy, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA*
[4]*School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA*

Accepted 2017 July 28. Received 2017 July 28; in original form 2016 November 19

ABSTRACT

We develop an automated technique for detecting damped Ly α absorbers (DLAs) along spectroscopic lines of sight to quasi-stellar objects (QSOs or quasars). The detection of DLAs in large-scale spectroscopic surveys such as SDSS III sheds light on galaxy formation at high redshift, showing the nucleation of galaxies from diffuse gas. We use nearly 50 000 QSO spectra to learn a novel tailored Gaussian process model for quasar emission spectra, which we apply to the DLA detection problem via Bayesian model selection. We propose models for identifying an arbitrary number of DLAs along a given line of sight. We demonstrate our method's effectiveness using a large-scale validation experiment, with excellent performance. We also provide a catalogue of our results applied to 162 858 spectra from SDSS-III data release 12.

Key words: methods: statistical – intergalactic medium – quasars: absorption lines – galaxies: statistics.

1 INTRODUCTION

The damped Ly α systems (DLAs; Wolfe et al. 1986; Wolfe, Gawiser & Prochaska 2005) define the class of absorption-line systems discovered in the rest-frame UV spectra of distant quasars, with H I column densities $N_{\mathrm{H\,I}} > 2 \times 10^{20}\,\mathrm{cm}^{-2}$, as measured from the analysis of damping wings in the Ly α profile. Recent spectroscopic quasar surveys such as the Sloan Digital Sky Survey (SDSS; York et al. 2000) have produced a vast sample of quasar spectra showing Ly α absorption at $z > 2$. SDSS III has measured nearly 300 000 quasar spectra over its brief history. Even larger surveys, such as the Dark Energy Spectroscopic Instrument (DESI[1]) survey, soon plan to observe 1–2 million quasars. Finding DLAs in these surveys has historically involved a combination of automated template fitting and visual inspection, but visual inspection is clearly infeasible with the size of upcoming data sets. Furthermore, SDSS data trades off low signal-to-noise ratios (SNRs) for statistical power, making detection of even distinctive signals such as DLAs substantially harder, and making noise-induced systematic error hard to control.

There have been several previous DLA searches in SDSS. These include a visual-inspection survey (Slosar et al. 2011), visually guided Voigt-profile fitting (Prochaska, Herbert-Fort & Wolfe 2005; Prochaska & Wolfe 2009) and two automated approaches: a template-matching approach (Noterdaeme et al. 2012), and an unpublished machine-learning approach using Fisher discriminant analysis (Carithers 2012). Although these methods have had some success in detecting large DLA catalogues, their reliance entirely on templates made them subject to hard-to-quantify systematic biases. In particular, these methods lack an explicit global model of quasar emission beyond simple continuum estimation, and the lack of such a model may give rise to unexpected false positives.

We present a new, completely automated method based on a rigorous Bayesian model-selection framework. We model the quasar spectra, including the continuum and non-DLA absorption, using Gaussian process (GP; Rasmussen & Williams 2006) models with a bespoke covariance function. Earlier catalogues are used as prior information to train the covariance. We provide a catalogue of our results on 162 858 QSOs with $z \geq 2.15$ from data release 12 of SDSS III, demonstrating that our method scales to very large data sets, making it ideally suited for future surveys. Furthermore, as our method relies on a well-defined probabilistic framework, it allows us to estimate the probability that each system is indeed a DLA, rather than a noise fluctuation, degrading gracefully for low-SNR observations. This property allows us to obtain substantially more-reliable measurements of the statistics of the DLA population in situations with reliable uncertainties even where systematic uncertainty dominates (Bird, Garnett & Ho 2017). We are also able to extend our catalogue to high redshift even with low-quality data.

Our method is applicable not just to DLAs, but also to other classes of absorption systems, such as Lyman limit systems and metal absorbers, which we intend to examine in future work. We focus on DLAs here both because of the large body of prior work which enables us to thoroughly verify our catalogues, and the intrinsic importance of these systems.

DLAs are a direct probe of neutral gas at densities close to those required to form stars (Cen 2012). The exact nature of the

★ E-mail: garnett@wustl.edu
[1] http://desi.lbl.gov/wp-content/uploads/2014/04/fdr-science-biblatex.pdf

systems hosting DLAs was initially debated, with kinematic data combined with simple semi-analytic models appearing to indicate objects similar in size to present day star-forming galaxies (Prochaska & Wolfe 1997; Jedamzik & Prochaska 1998; Maller et al. 2001), whereas early simulations produced clumps closer in size to dwarf galaxies (Haehnelt, Steinmetz & Rauch 1998; Okoshi & Nagashima 2005). Recent numerical simulations are able to reproduce most observations with neutral hydrogen clouds stretching almost to the virial radius of objects larger than dwarfs, but smaller than present day star-forming galaxies (Pontzen et al. 2008; Rahmati et al. 2013; Bird et al. 2015). Associated galactic stellar components have been detected in a few, particularly neutral hydrogen and metal-rich systems at low redshift (Le Brun et al. 1997; Rao et al. 2003; Chen 2005). However, unbiased surveys have placed strong upper limits on the star-formation rates of the median DLA (Fumagalli et al. 2015), indicating that DLAs are associated with low star-formation rate objects.

DLAs represent our only probe of small- to moderate-sized galaxies at high redshift, and are known to have dominated the neutral-gas content of the Universe from redshift $z = 5$ (when the Universe was 1.2 Gyr old) to today (Gardner et al. 1997; Wolfe et al. 2005). The neutral gas in these systems ultimately accretes on to galactic haloes and fuels star formation. Thus, their abundance as a function of redshift provides strong constraints on models of galaxy formation (Bird et al. 2014). Our work, including publicly available software, will not only provide observers with a new automated tool for detecting these objects, but also provide theorists with a reliable catalogue on which to base theoretical models.

2 NOTATION

We will briefly establish some notation. Consider a QSO with redshift $z_{\rm QSO}$; we will always assume that $z_{\rm QSO}$ is known, allowing us to work in the quasar rest frame. We will notate a QSO's true emission spectrum by a function $f : \mathbb{R} \to \mathbb{R}$, where $f(\lambda)$ represents the flux corresponding to rest wavelength λ. Without subscript, λ will always refer to quasar rest wavelengths, $\lambda_{\rm rest}$, rather than observed wavelengths, $\lambda_{\rm obs}$. Note that the spectral emission function f is never directly observed, both due to measurement error and due to absorption by intervening matter along the line of sight. We will denote the observed flux by a corresponding function $y(\lambda)$, which will again be a function of the rest wavelengths.

Spectrographic observations of a QSO are made at a discrete set of wavelengths $\boldsymbol{\lambda}$, for which we observe a corresponding vector of flux measurements $\boldsymbol{y}$, where we have defined $y_i = y(\lambda_i)$. For a given QSO, we will represent the set of observation locations and values $(\boldsymbol{\lambda}, \boldsymbol{y})$ by $\mathcal{D}$.

We will often encounter data with missing values due to observation-dependent pixel masking. When required, we will represent these in the text with a special value called NaN (for 'not a number'). Calculations on data containing NaN s will always ignore these values.

3 BAYESIAN MODEL SELECTION

Our approach to DLA detection will depend on *Bayesian model selection,* which will allow us to directly compute the probability that a given quasar sightline contains a DLA. We will develop two probabilistic models for a given set of spectroscopic observations $\mathcal{D}$: one for sightlines with intervening DLAs and one for those without. Then, given the available data, we will compute the posterior probability that the former model is correct. We will give a high-level overview of Bayesian model selection below, then proceed to describe our models for DLA detection below.

Let $\mathcal{M}$ be a probabilistic model, and let θ represent a vector of parameters for this model (if any). Given a set of observed data $\mathcal{D}$ and a set of candidate models $\{\mathcal{M}_i\}$ containing $\mathcal{M}$, we wish to compute the probability of $\mathcal{M}$ being the correct model to explain $\mathcal{D}$. The key quantity of interest to model selection is the so-called *model evidence*:

$$p(\mathcal{D} \mid \mathcal{M}) = \int p(\mathcal{D} \mid \mathcal{M}, \theta) p(\theta \mid \mathcal{M}) \, \mathrm{d}\theta, \tag{1}$$

which represents the probability of having generating the observed data with the model, after having integrated out any uncertainty in the parameter vector θ. Given the model evidence, we can apply Bayes' rule to compute the posterior probability of the model given the data:

$$\Pr(\mathcal{M} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathcal{M}) \Pr(\mathcal{M})}{p(\mathcal{D})} = \frac{p(\mathcal{D} \mid \mathcal{M}) \Pr(\mathcal{M})}{\sum_i p(\mathcal{D} \mid \mathcal{M}_i) \Pr(\mathcal{M}_i)}, \tag{2}$$

where $\Pr(\mathcal{M})$ represents the prior probability of the model. Notice that computing the posterior probability of $\mathcal{M}$ requires computing the normalizing constant in the denominator.

We will develop two models for spectroscopic observations of QSOs, $\mathcal{M}_{\neg \rm DLA}$, for lines of sight that do not contain intervening DLAs, and $\mathcal{M}_{\rm DLA}$, for those that do. Both of these models will rely heavily on GPs, which we will introduce below.

4 GAUSSIAN PROCESSES

The main object of interest we wish to perform inference about is a given QSO's emission function $f(\lambda)$. This is in general a complicated function with no simple parametric form available, so we will instead use nonparametric inference techniques to reason about it. GPs provide a powerful nonparametric framework for modelling unknown functions, which we will adopt. See Rasmussen & Williams (2006) for an extensive introduction to GPs.

4.1 Definition and prior distribution

Let $\mathcal{X}$ be an arbitrary input space, for example the real line $\mathbb{R}$, and let $f : \mathcal{X} \to \mathbb{R}$ be a real-valued function on $\mathcal{X}$ we wish to model. We will continue to use λ to indicate inputs to the function f. A GP is an extension of the multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ to infinite domains. Like the multivariate Gaussian distribution, a GP is fully specified by its first two central moments: a mean function $\mu(\lambda)$ and a positive semidefinite covariance function $K(\lambda, \lambda')$:[2]

$$\mu(\lambda) = \mathbb{E}[f(\lambda) \mid \lambda];$$
$$K(\lambda, \lambda') = \mathrm{cov}[f(\lambda), f(\lambda') \mid \lambda, \lambda'].$$

The former describes the pointwise expected value of the function and the latter describes the correlation around the mean. Given μ and K, we may endow the function space f with a GP prior probability distribution:

$$p(f) = \mathcal{GP}(f; \mu, K). \tag{3}$$

[2] A function $K : \mathcal{X}^2 \to \mathbb{R}$ is *positive semidefinite* if, for every finite subset $\Lambda = \{\lambda_i\}_{i=1}^n \subset \mathcal{X}$, the $n \times n$ *Gram matrix* $\mathbf{A}$, defined by $A_{ij} = K(\lambda_i, \lambda_j)$, satisfies $\boldsymbol{c}^\top \mathbf{A} \boldsymbol{c} \geq 0$ for all $\boldsymbol{c} \in \mathbb{R}^n$.

The defining characteristic of a GP is that given a finite set of inputs $\boldsymbol{\lambda}$, the corresponding vector of function values $\boldsymbol{f} = f(\boldsymbol{\lambda})$ is multivariate Gaussian distributed:

$$p(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}; \mu(\boldsymbol{\lambda}), K(\boldsymbol{\lambda}, \boldsymbol{\lambda})), \tag{4}$$

where the mean vector and covariance matrix are derived simply by evaluating the mean and covariance functions at the inputs $\boldsymbol{\lambda}$, and the multivariate Gaussian probability distribution function is given by

$$\mathcal{N}(\boldsymbol{f}; \boldsymbol{\mu}, \mathbf{K}) = \frac{1}{\sqrt{(2\pi)^d \det \mathbf{K}}} \exp\left(-\frac{1}{2}(\boldsymbol{f} - \boldsymbol{\mu})^\top \mathbf{K}^{-1} (\boldsymbol{f} - \boldsymbol{\mu})\right), \tag{5}$$

where d is the dimension of $\boldsymbol{f}$.

4.2 Observation model

Consider a set of noisy observations $\mathcal{D} = (\boldsymbol{\lambda}, \boldsymbol{y})$ made at input locations $\boldsymbol{\lambda}$. Our GP prior on f implies a multivariate Gaussian distribution for the corresponding (unknown, so-called *latent*) function values $\boldsymbol{f} = f(\boldsymbol{\lambda})$, but does not specify the relationship between these values and our observations $\boldsymbol{y}$. Instead, we must further model the mechanism generating our observations, which we will encode by a distribution

$$p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{f}). \tag{6}$$

In general, this can be any arbitrary probabilistic model, but here we will assume additive Gaussian noise.

Given a single input location λ, we assume that the corresponding observed value y is realized by corrupting the true value of the latent function $f(\lambda)$ by zero-mean additive Gaussian noise with known variance $\sigma(\lambda)^2$:

$$p(y \mid \lambda, f(\lambda), \sigma(\lambda)) = \mathcal{N}(y; f(\lambda), \sigma(\lambda)^2). \tag{7}$$

We assume the noise process is independent for every λ, but note that we do not make a homoskedasticity assumption; rather, we allow the noise variance to depend on λ. This capability to handle heteroskedastic noise is critical for the analysis of spectroscopic measurements, where the noise associated with flux measurements can vary widely as a function of wavelength.

Returning to our entire set of observations $\mathcal{D} = (\boldsymbol{\lambda}, \boldsymbol{y})$, we assume that the noise variance associated with each of these measurements is known and given by a corresponding vector $\boldsymbol{v}$, with $v_i = \sigma(\lambda_i)^2$. Given our model for individual observations (7) and the noise independence assumption, the entire observation model is given by

$$p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{f}, \boldsymbol{v}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{f}, \mathbf{V}), \tag{8}$$

where $\mathbf{V} = \mathrm{diag}\boldsymbol{v}$, and we use the diag notation applied to a vector to refer to a square diagonal matrix with leading diagonal equal to the specified vector.

4.2.1 *Prior of noisy observations*

Given a set of observations locations $\boldsymbol{\lambda}$ and a corresponding vector of noise variances $\boldsymbol{v}$, we may use the above to compute the prior distribution for a corresponding vector of observations $\boldsymbol{y}$ by marginalizing the latent function values $\boldsymbol{f}$:

$$\begin{aligned} p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{v}) &= \int p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{f}, \boldsymbol{v}) p(\boldsymbol{f} \mid \boldsymbol{\lambda}) \, \mathrm{d}\boldsymbol{f} \\ &= \int \mathcal{N}(\boldsymbol{y}; \boldsymbol{f}, \mathbf{V}) \, \mathcal{N}(\boldsymbol{f}; \mu(\boldsymbol{\lambda}), K(\boldsymbol{\lambda}, \boldsymbol{\lambda})) \, \mathrm{d}\boldsymbol{f} \\ &= \mathcal{N}(\boldsymbol{y}; \mu(\boldsymbol{\lambda}), K(\boldsymbol{\lambda}, \boldsymbol{\lambda}) + \mathbf{V}), \end{aligned} \tag{9}$$

where we have used the fact that Gaussians are closed under convolution to compute the integral in closed form.

In typical applications of GP inference, the prior mean function μ and prior covariance function K would be selected from numerous several off-the-shelf solutions available for this purpose; however, none of these would be directly appropriate for modelling QSO emission spectra, due to their somewhat complex nature. Typical parametric covariance functions, for example, tend to be translation invariant and encode strictly decreasing covariance as a function of the distance between inputs.[3] QSO emission spectra, however, are neither stationary, nor should we expect the covariance to be diagonal dominant. For example, strong off-diagonal correlations must exist between potentially distant emission lines, such as members of the Lyman series. Rather, below we will construct a custom GP prior distribution for modelling these spectra in the next section.

5 LEARNING A GP PRIOR FOR QSO SPECTRA

We wish to construct a GP prior for QSO spectra, specifically, those that do not contain an intervening DLA along the line of sight. This will form the basis for our null model $\mathcal{M}_{\neg\mathrm{DLA}}$. We will later extend this to form our DLA model $\mathcal{M}_{\mathrm{DLA}}$.

As described in the previous section, a GP is defined entirely by its first two moments: a mean function $\mu(\lambda)$ and a covariance function $K(\lambda, \lambda')$. Therefore, our goal in this section will be to derive reasonable prior choices for these functions. Due to the complex structure of QSO emission spectra, our approach will be to make as few assumptions as possible. Instead, we adopt a data-driven approach and learn an appropriate model given over 48 000 examples contained in a previously compiled catalogue of quasar spectra recorded by the BOSS spectrograph (Smee et al. 2013).

5.1 Data

Together, SDSS I, II (Abazajian et al. 2009) and III (Eisenstein et al. 2011) used a drift-scanning mosaic CCD camera (Gunn et al. 1998) to image over one-third of the sky (14 555 square degrees) in five photometric bandpasses (Fukugita et al. 1996; Smith et al. 2002; Doi et al. 2010) to a limiting magnitude of $r < 22.5$ using the dedicated 2.5-m Sloan Telescope (Gunn et al. 2006) located at Apache Point Observatory in New Mexico.

The Baryon Oscillation Spectroscopic Survey (BOSS), a part of the SDSS-III survey (Eisenstein et al. 2011), has obtained spectra of 1.5 million galaxies approximately volume limited out to $z \sim 0.6$ (Reid et al. 2016), and an additional 150 000 spectra of high-redshift quasars and ancillary sources. BOSS has measured the characteristic scale imprinted by baryon acoustic oscillations (BAOs) in the early Universe from the spatial distribution of galaxies at $z \sim 0.5$ and the H I absorption lines in the intergalactic medium at $z \sim 2.3$ (Anderson et al. 2012, 2014; Aubourg et al. 2015). The quasar target selection is described in Bovy et al. (2011) and Ross et al. (2012). Here, we use data included in data releases 9 (DR9; Ahn et al. 2012) and data releases 12 (DR12; Ahn et al. 2014) of SDSS III; in particular, we primarily use the associated quasar catalogues from various data releases[4] (Pâris et al. 2012, 2014).

[3] The Wiener process, modelling the sample paths of Brownian motion, is a GP with such a covariance function.

[4] http://www.sdss.org/dr12/algorithms/boss-dr12-quasar-catalog/

5.1.1 Description of data

We used the QSO spectra from the BOSS DR9 Ly α forest sample (Lee et al. 2013) to train our GP model. This sample comprises 54 468 QSO spectra with $z_{\text{QSO}} > 2.15$ from the DR9 release appropriate for Ly α forest analysis. An analogous model built from the entire DR12 sample will be published along with manuscript for general-purpose use, along with the source code (in MATLAB) we used to train our model and conduct our investigation.

The Ly α forest sample was augmented with a previously compiled 'concordance' DLA catalogue (Carithers 2012), combining the results of three previous DLA searches. These include a visual-inspection survey (Slosar et al. 2011) and two previous automated approaches: a template-matching approach (Noterdaeme et al. 2012), and an unpublished machine-learning approach using Fisher discriminant analysis (Carithers 2012). Any line of sight flagged in at least two of these catalogues as containing a DLA is included in the concordance catalogue. Both previous automated DLA searches also produced estimates of the absorber redshift z_{DLA} and column density $\log_{10} N_{\text{H I}}$. The concordance catalogue also includes these estimates for flagged sightlines; when a sightline is included in both automated catalogues, the arithmetic mean of the associated estimates was recorded. A total of 5854 lines of sight are flagged as containing an intervening DLA in the catalogue (10.7 per cent).

5.2 Modelling decisions

To avoid effects due to redshift, we will build our emission model for wavelengths in the rest frame of the QSO. Furthermore, to account for arbitrary scaling of flux measurements, we will build a GP prior for normalized flux. Specifically, given the observed flux of a QSO, we normalize all flux measurements by dividing by the median flux observed between 1310 Å and 1325 Å in the rest frame of the QSO, a region redwards of the Ly α forest and void of major emission features.

Because this study is concerned with identifying DLAs, we will only model the flux bluewards of the Ly α emission in the rest frame of a given QSO.[5] Specifically, we model emissions in the range spanning from the Lyman limit to the Ly α line in the QSO rest frame.[6] Our approach will be to learn a mean vector and covariance matrix on a dense grid of wavelengths in this range, which we will then interpolate as required by a particular set of observed wavelengths.[7] The chosen grid was the set of wavelengths

$$\lambda \in [911.75 \text{ Å}, 1215.75 \text{ Å}], \tag{10}$$

with a linearly equal spacing of $\Delta\lambda = 0.25$ Å.[8] This resulted in a vector of input locations $\boldsymbol{\lambda}$ with $|\boldsymbol{\lambda}| = N_{\text{pixels}} = 1\,217$ pixels.

[5] One could consider an extension of our approach where metal absorption lines corresponding to wavelengths redwards of Ly α were considered, requiring modelling spectra over a larger range of wavelengths; however, we will not do so here.

[6] We stop at the Lyman limit to avoid being confused by the Lyman break associated with Lyman limit systems.

[7] Such interpolation introduces minor correlation between pixels; however, this effect is unlikely to be large.

[8] This represents about 3–4 times the maximum pixel separation of the BOSS spectrograph; the minimum separation in a single BOSS spectrum's measured wavelengths is approximately $(10^{\log_{10} 3600 + 0.0001} - 3600)$ Å $\approx$ 0.83 Å. Note, however, that we have tens of thousands of observations corresponding to each of the wavelengths in our chosen grid.

Given a GP prior for QSO emission spectra, $p(f) = \mathcal{GP}(f; \mu, K)$, the prior distribution for emissions on the chosen grid $\boldsymbol{\lambda}$, $\boldsymbol{f} = f(\boldsymbol{\lambda})$ is a multivariate Gaussian:

$$p(\boldsymbol{f} \mid \boldsymbol{\lambda}, z_{\text{QSO}}) = \mathcal{N}(\boldsymbol{f}; \boldsymbol{\mu}, \mathbf{K}), \tag{11}$$

where $\boldsymbol{\mu} = \mu(\boldsymbol{\lambda})$ and $\mathbf{K} = K(\boldsymbol{\lambda}, \boldsymbol{\lambda})$. Note that we must condition on the QSO redshift z_{QSO} because it is required for shifting into the quasar rest frame.

As mentioned previously, however, we can never observe f directly, due to both measurement error and absorption by intervening matter along the line of sight. The former can be handled easily for our spectra by using the pipeline error estimates in the role of the noise vector $\boldsymbol{v}$ (see Section 4.2). However, the latter is more problematic, especially in our chosen region, which includes the Ly α forest. In principle, if we knew the exact nature of the intervening matter, we could model this absorption explicitly; however, this is unrealistic. We will instead model the effect of small absorption phenomena (absorption by objects with column density below the DLA limit, $\log_{10} N_{\text{H I}} < 20.3$) by an additional additive wavelength- and redshift-dependent Gaussian noise term, which we will learn. Therefore, the characteristic 'dips' of the Ly α forest will be modelled as noisy deviations from the true underlying smooth continuum. Later, we will explicitly model larger absorption phenomena (DLAs with $\log_{10} N_{\text{H I}} \geq 20.3$) to build our DLA model $\mathcal{M}_{\text{DLA}}$.

The mathematical consequence of this modelling decision is as follows. Consider the arbitrary GP model in equation (11). We wish to model the associated spectroscopic observation values on the chosen grid, $\boldsymbol{y} = y(\boldsymbol{\lambda})$. Suppose that the measurement noise vector $\boldsymbol{v} = \sigma(\boldsymbol{\lambda})^2$ has been specified. During our exposition on GPs, we described the additive Gaussian noise observation model (8). The model we adapt here will involve a shared non-DLA absorption 'noise' vector $\boldsymbol{\omega}$, defined in the quasar rest frame, modelling absorption deviations from the QSO continuum.

Due to the evolution of the Ly α forest flux with redshift, we additionally incorporate a simple power-law redshift dependence into this absorption noise model. Namely, the absorption noise standard deviation we incorporate at an observed wavelength λ_{obs} is defined to be

$$\omega'(\lambda_{\text{obs}}, \lambda_{\text{rest}}) = \omega(\lambda_{\text{rest}}) s(z(\lambda_{\text{obs}}))^2; \tag{12}$$

$$s(z) = 1 - \exp(-\tau_0 (1+z)^{\beta}) + c_0, \tag{13}$$

where $\omega(\lambda_{\text{rest}})$ is the shared absorption noise corresponding to the wavelength in the quasar rest frame, c_0, τ_0 and β are constants, and $z(\lambda_{\text{obs}})$ is the redshift of Ly α at the observed wavelength. Hence, our model depends on the redshift of the quasar as well as the redshift of Ly α along the line of sight.

The resulting observation model is

$$p(\boldsymbol{y} \mid \boldsymbol{f}, \boldsymbol{v}, \boldsymbol{\omega}, z_{\text{QSO}}, \mathcal{M}_{\neg\text{DLA}}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{f}, \boldsymbol{\Omega} + \mathbf{V}), \tag{14}$$

where $\boldsymbol{\Omega} = \text{diag}\boldsymbol{\omega}'$, and $\boldsymbol{\omega}'$ incorporates the redshift dependence as defined above. Therefore, given our chosen grid $\boldsymbol{\lambda}$, the prior distribution of associated spectroscopic observations $\boldsymbol{y}$ is

$$p(\boldsymbol{y} \mid \boldsymbol{v}, \boldsymbol{\omega}, z_{\text{QSO}}, \mathcal{M}_{\neg\text{DLA}}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}, \mathbf{K} + \boldsymbol{\Omega} + \mathbf{V}), \tag{15}$$

derived analogously to equation (9). Our goal now is to learn appropriate values for $\boldsymbol{\mu}$, $\boldsymbol{K}$, $\boldsymbol{\omega}$, c_0, τ_0 and β, which will fully specify our null model $\mathcal{M}_{\neg\text{DLA}}$.

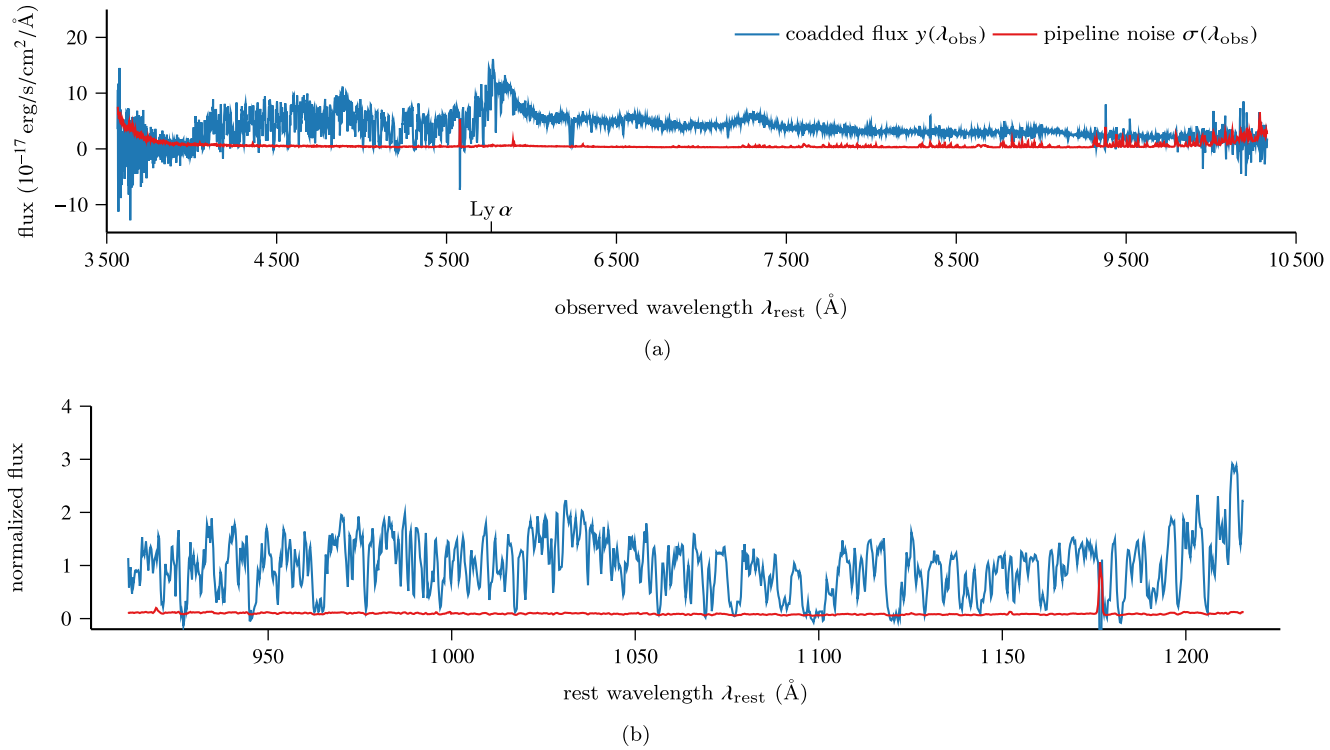


Figure 1. An illustration of the data preprocessing procedure for object SDSS 020712.80+052753.4, (plate, MJD, fibre) = (4401, 55510, 338); $z_{\rm QSO} = 3.741$. This QSO is included in the DLA concordance catalogue with $(z_{\rm DLA}, \log_{10} N_{\rm H\,I}) = (3.283, 20.39)$, corresponding to central absorption wavelength $\lambda_{\rm obs} = 5\,206$ Å or $\lambda_{\rm rest} = 1\,098$ Å in the QSO rest frame. The wavelengths are shifted to the QSO rest frame and pixels outside $\lambda_{\rm rest} \in [911.75\text{ Å}, 1215.75\text{ Å}]$ are discarded. Finally, the flux and noise estimates are normalized by dividing by the median flux observed in the range of [1310 Å, 1325 Å]. The final result is shown in (b).

5.3 Learning appropriate parameters

To build our null model, we took the $N_{\rm spec} = 48\,614$ spectra from the BOSS DR9 Ly α forest sample that are putatively absent of intervening DLAs. We prepared each of these spectra for processing in an identical manner as follows:

(i) The augmented spectrum file was loaded and the (wavelength, observed flux, pipeline noise variance) = (λ, y, v) measurements in the chosen modelled region were extracted.
(ii) The wavelengths were shifted to the rest frame of the QSO.
(iii) Flux measurements with serious pixel mask bit flags (`FULL-REJECT`, `NOSKY`, `BRIGHTSKY`, `NODATA`) set by the SDSS pipeline were masked (replaced by NaN).
(iv) The flux normalizer was determined by examining the region corresponding to [1310, 1325] Å in the rest frame of the quasar; the median nonmasked value in this range was used for normalization.
(v) The flux and noise variance were normalized with the value computed in the last step.

Finally, we linearly interpolated the resulting flux and noise variance measurements of each spectrum on to the chosen wavelength grid $\boldsymbol{\lambda}$. Note that this interpolation preserved NaN s; we did not 'interpolate through' masked pixels. We also did not extrapolate beyond the range of wavelengths present in each spectrum. The preprocessing procedure is illustrated in Fig. 1 on a spectrum we will use as a running example.

We collect the resulting interpolated vectors into $(N_{\rm spec} \times N_{\rm pixels})$ matrices $\mathbf{Y}$ and $\mathbf{V}$, containing the normalized flux and noise variance vectors, respectively. For QSO i, we will write $\boldsymbol{y}_i$ and $\boldsymbol{v}_i$ to represent the corresponding observed flux and noise variance vectors and will define $\mathbf{V}_i = \mathrm{diag}\boldsymbol{v}_i$.

Due to masked pixels and varying redshifts of each QSO, the $\mathbf{Y}$ and $\mathbf{V}$ matrices contain numerous missing values, especially on the blue end. Fig. 2 shows the portion of available data as a function of wavelength.

5.3.1 *Learning the mean*

Identifying an appropriate mean vector $\boldsymbol{\mu}$ is straightforward with so many example spectra. We simply found the mean recorded value for each rest wavelength in our grid across the available measurements:

$$\mu_j = \frac{1}{N_{\neg\mathrm{NaN}}} \sum_{y_{ij} \neq \mathrm{NaN}} y_{ij}. \tag{16}$$

Note that the sample mean is the maximum-likelihood estimator for $\boldsymbol{\mu}$. The learned mean vector $\boldsymbol{\mu}$ is plotted in Fig. 3. Several emission features are obvious.

5.3.2 *Learning the flux covariance and additional absorption noise*

We will use standard unconstrained optimization techniques to learn the covariance matrix $\mathbf{K}$ and absorption 'noise' vector $\boldsymbol{\omega}$. Without further structural assumptions on $\mathbf{K}$, however, we would be forced to learn $N^2_{\rm pixels} \approx 1.5 \times 10^6$ entries. Instead, we will use a low-rank decomposition to limit the number of free variables in our model:

$$\mathbf{K} = \mathbf{M}\mathbf{M}^\top, \tag{17}$$

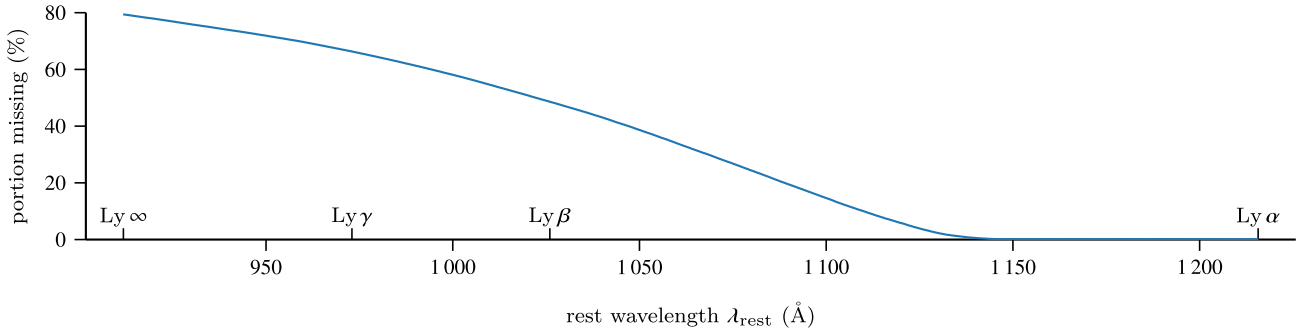


Figure 2. The portion of missing pixels as a function of wavelength for the 48 614 QSOs in the BOSS DR9 Ly α forest sample used for learning our GP model.

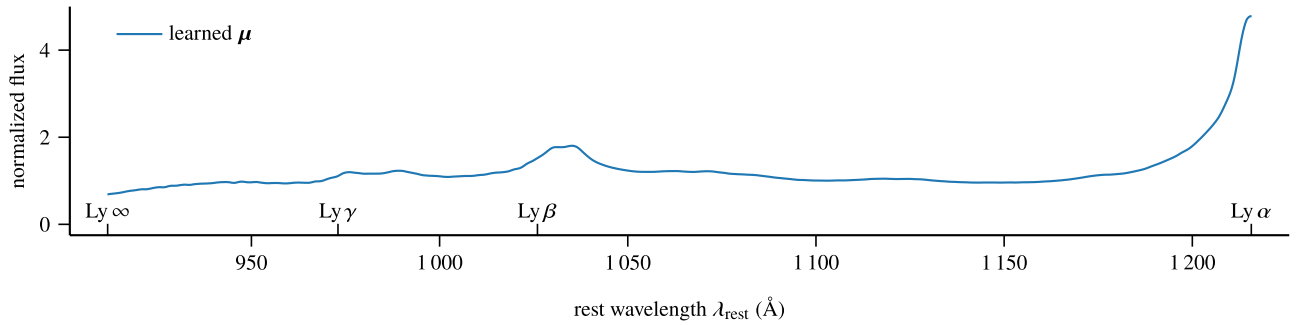


Figure 3. The learned mean vector $\boldsymbol{\mu}$ derived by taking the median across the stacked spectra. The vector has been smoothed with a 4-pixel (1 Å) boxcar function for clarity on the blue end.

where $\mathbf{M}$ is an $(N_{\text{pixels}} \times k)$ matrix with $k \ll N_{\text{pixels}}$. This decomposition guarantees that $\mathbf{K}$ will be positive semidefinite (and thus a valid covariance matrix) for any $\mathbf{M}$. Note that this decomposition is similar to that encountered in principal component analysis (PCA); however, note that we do not constrain the columns of $\mathbf{M}$ (the 'eigenspectra') to be orthogonal. Here, we took $k = 20$.

We assume that each of our measured flux vectors is an independent realization drawn from a common observation prior (15):

$$p(\mathbf{Y} \mid \boldsymbol{\lambda}, \mathbf{V}, \mathbf{M}, \boldsymbol{\omega}, \boldsymbol{z}_{\text{QSO}}, \mathcal{M}_{\neg\text{DLA}}) = \prod_{i=1}^{N_{\text{spec}}} \mathcal{N}(\boldsymbol{y}_i; \boldsymbol{\mu}, \mathbf{K} + \boldsymbol{\Omega} + \mathbf{V}_i), \tag{18}$$

where $\boldsymbol{z}_{\text{QSO}}$ is a vector concatenating the redshifts of the quasars, and all NaN values are ignored in the computation. That is, in the ith entry of the product, we only use the entries of $\boldsymbol{\mu}$, $\boldsymbol{v}_i$ and $\boldsymbol{w}$', and only the rows of $\mathbf{M}$, corresponding to the nonmasked values in $\boldsymbol{y}_i$.

We define the log likelihood of the data, $\mathcal{L}$, as a function of the covariance parameters $\mathbf{M}$ and $\boldsymbol{\omega}$. To simplify the notation, we first define the following quantities:

$$\boldsymbol{\Sigma}_i = \mathbf{K} + \boldsymbol{\Omega} + \mathbf{V}_i; \tag{19}$$

$$\boldsymbol{\alpha}_i = \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}). \tag{20}$$

Now the log likelihood is

$$\begin{aligned}\mathcal{L}(\mathbf{M}, \boldsymbol{\omega}) &= \log p(\mathbf{Y} \mid \boldsymbol{\lambda}, \mathbf{V}, \mathbf{M}, \boldsymbol{\omega}, \boldsymbol{z}_{\text{QSO}}, \mathcal{M}_{\neg\text{DLA}}) \\ &= \sum_{i=1}^{N_{\text{spec}}} \log \mathcal{N}(\boldsymbol{y}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}_i) \\ &= \sum_{i=1}^{N_{\text{spec}}} -\frac{1}{2}\left(\boldsymbol{\alpha}_i^\top(\boldsymbol{y}_i - \boldsymbol{\mu}) + \log\det\boldsymbol{\Sigma}_i + N_i \log 2\pi\right), \end{aligned} \tag{21}$$

where N_i is the number of non-NaN pixels in $\boldsymbol{y}_i$. We will maximize $\mathcal{L}(\mathbf{M}, \boldsymbol{\omega})$ with respect to the covariance parameters to derive our model, giving the emission model most likely to have generated our data. To enable unconstrained optimization, we parametrize the $\boldsymbol{\omega}$ parameter by its natural logarithm, guaranteeing every entry of $\boldsymbol{\omega}$ is positive after exponentiation. In the context of its role in our model, this is equivalent to reasoning about the optical depth τ rather than the absorption $\exp(-\tau)$.

An important feature of our particular choice of model is that we can compute the matrix inverse and the log determinant of $(\mathbf{K} + \boldsymbol{\Omega} + \mathbf{V})$ quickly. Namely, this matrix has the form $\mathbf{M}\mathbf{M}^\top + \mathbf{D}$, where $\mathbf{D}$ is diagonal. We may apply the Woodbury identity to derive

$$(\mathbf{M}\mathbf{M}^\top + \mathbf{D})^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{M}(\mathbf{I} + \mathbf{M}^\top\mathbf{D}^{-1}\mathbf{M})^{-1}\mathbf{M}^\top\mathbf{D}^{-1}, \tag{22}$$

where $\mathbf{I}$ is the identity matrix. Note the nominally $N_{\text{pixels}} \times N_{\text{pixels}}$ inverse can be computed via a much less expensive $k \times k$ inverse. Similarly, we may use the Sylvester determinant theorem to derive

$$\log\det(\mathbf{M}\mathbf{M}^\top + \mathbf{D}) = \log\det\mathbf{D} + \log\det(\mathbf{I} + \mathbf{M}^\top\mathbf{D}^{-1}\mathbf{M}), \tag{23}$$

again reducing the problem to a determinant on a $k \times k$ matrix.

To maximize our joint log likelihood, we applied the L-BFGS algorithm, a quasi-Newton algorithm for unconstrained optimization. The required partial derivatives are

$$\frac{\partial\mathcal{L}_i}{\partial\mathbf{M}} = \left(\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^\top - \boldsymbol{\Sigma}_i^{-1}\right)\mathbf{M}; \tag{24}$$

$$\frac{\partial\mathcal{L}_i}{\partial\log\boldsymbol{\omega}} = \boldsymbol{\omega}' \circ \left(\boldsymbol{\alpha}_i^2 - \operatorname{diag}\boldsymbol{\Sigma}_i^{-1}\right), \tag{25}$$

where $\circ$ is the Hadamard (element-wise) product, and we define diag applied to a square matrix to return its leading diagonal as a vector. The partial derivatives with respect to $\log c_0$, $\log\tau_0$ and

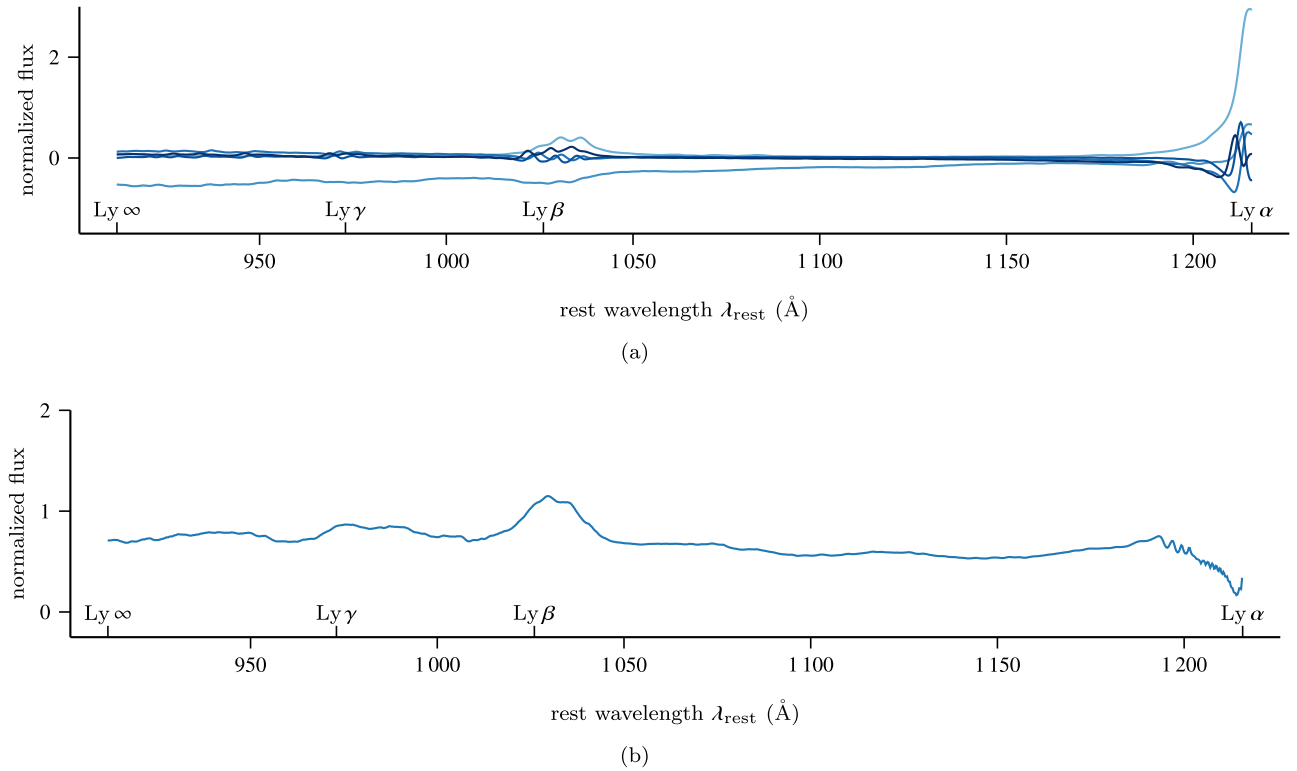


Figure 4. (a) The first five columns of the learned M and (b) the learned absorption noise vector $\boldsymbol{\omega}$, both learned from the 48 614 QSOs in the BOSS DR9 Ly α forest sample. Both have been smoothed with a 4-pixel (1 Å) boxcar function for clarity on the blue end.

$\log \beta$ all have the same form:

$$\frac{\partial \mathcal{L}_i}{\partial \log x} = \boldsymbol{\alpha}_i^\top \left(\operatorname{diag} \frac{\partial \boldsymbol{\omega}'}{\partial x} \right) \boldsymbol{\alpha}_i + \left(\frac{\partial \boldsymbol{\omega}'}{\partial x} \right)^\top \operatorname{diag} \boldsymbol{\Sigma}^{-1}; \tag{26}$$

$$\frac{\partial \boldsymbol{\omega}'}{\partial \log c_0} = c_0 \, \boldsymbol{\omega} \circ s(\boldsymbol{z}); \tag{27}$$

$$\frac{\partial \boldsymbol{\omega}'}{\partial \log \tau_0} = \tau_0 \, \boldsymbol{\omega} \circ s(\boldsymbol{z}) \circ (1+\boldsymbol{z})^\beta \circ \exp(-\tau_0 (1+\boldsymbol{z})^\beta); \tag{28}$$

$$\frac{\partial \boldsymbol{\omega}'}{\partial \log \beta} = \beta \log \boldsymbol{z} \circ \frac{\partial \boldsymbol{\omega}'}{\partial \log \tau_0}, \tag{29}$$

where $\boldsymbol{z}$ is a vector of the Ly α redshifts corresponding to the observations, and the redshift contribution s is defined in equation (13).

We learned the decomposed covariance matrix $\mathbf{M}$, $\boldsymbol{\omega}$, c_0, τ_0 and β via L-BFGS on the selected training spectra. For this model learning phase only, we masked all pixels with noise variance larger than unity after normalization (i.e. pixels with SNRs below approximately 1). Note that these pixels were only masked here and at no other point in this study. The initial value for $\mathbf{M}$ was taken to be the top-20 principal components of $\mathbf{Y}$, estimated entry-wise using available data. Masking low-SNR pixels was required here because PCA, in its most basic form, does not account for noise in measured values, and our heteroskedastic noise is especially troublesome. The initial value of each entry in $\boldsymbol{\omega}$ was taken to be the sample variance of the corresponding column of $\mathbf{Y}$, ignoring NaN s.

The first five columns of the learned $\mathbf{M}$ and the learned absorption noise vector $\boldsymbol{\omega}$ are shown in Fig. 4. The corresponding covariance matrix $\mathbf{M}\mathbf{M}^\top$ is shown in Fig. 5. Features corresponding to the Lyman series are clearly visible, including strong off-diagonal correlations between pairs of emission lines. At least seven members of the Lyman series can be identified in the covariance entries corresponding to Ly α emission. This complex (and physically correct) structure was automatically learned from the data. The parameters for the redshift-dependent component of the absorption noise vector were

$$c_0 = 0.3371; \qquad \tau_0 = 0.01178; \qquad \beta = 1.797. \tag{30}$$

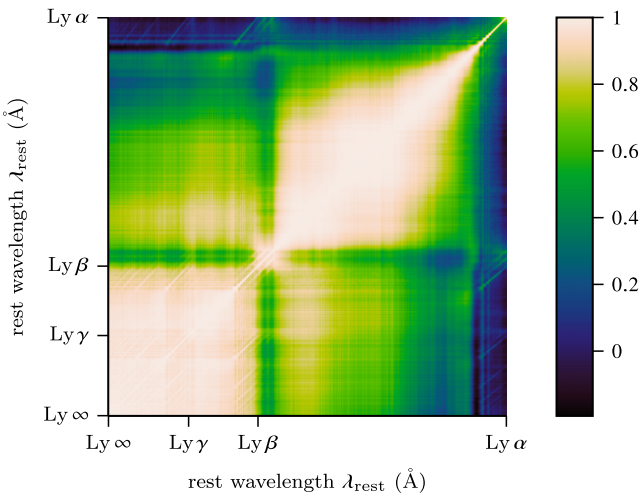


Figure 5. The observation covariance matrix $\mathbf{K}$ corresponding to the learned parameters shown in Fig. 4. The entries have been normalized to give unit diagonal; the entries are therefore correlations rather than raw covariances.

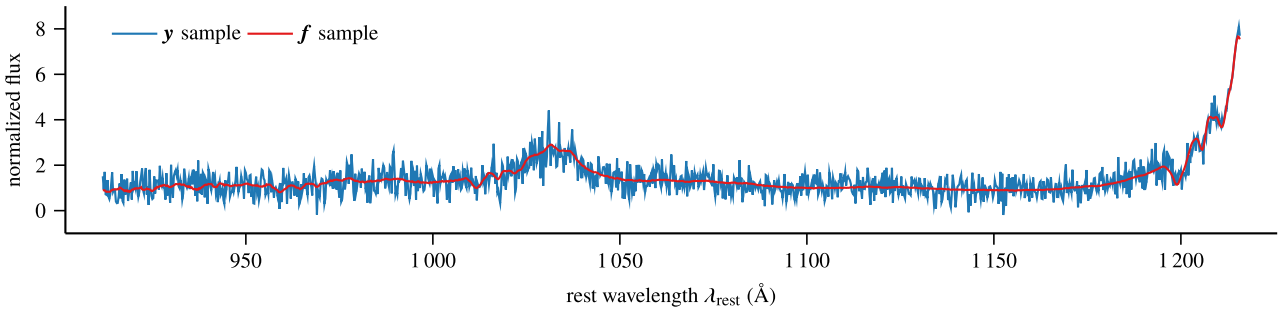


Figure 6. An example sample from our learned QSO emission spectrum model $\mathcal{GP}(\boldsymbol{f}; \boldsymbol{\mu}, \mathbf{K})$ (in red), and the corresponding sample after incorporating our additional absorption correction into the model, a draw from $p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{v}, \mathcal{M}_{\neg\mathrm{DLA}}) = \mathcal{GP}(\boldsymbol{y}; \boldsymbol{\mu}, \mathbf{K} + \boldsymbol{\Omega} + \mathbf{V})$ (in blue). Constant observation noise with variance $v = 0.1^2$ was simulated for the $\boldsymbol{y}$ sample.

We have now fully specified our GP prior for QSO emission spectra in the range of $\lambda \in [911.75\,\text{Å}, 1215.75\,\text{Å}]$. Fig. 6 demonstrates our model by showing an example sample from the prior distribution on QSO continuua $\boldsymbol{f}$, as well as a corresponding sample from the prior distribution on observations $\boldsymbol{y}$ incorporating our absorption 'noise' vector $\boldsymbol{\omega}$. The samples closely resemble actual observations.

Note that to apply our model to observations corresponding to a set of input wavelengths differing from the grid we used to learn the model, we simply interpolate (linearly) the learned $\boldsymbol{\mu}$, $\mathbf{K}$, and $\boldsymbol{\omega}$ on to the desired wavelengths. We may also account for redshift trivially should we wish to work with observed rather than rest wavelengths.

5.4 Model evidence

We note that our null model $\mathcal{M}_{\neg\mathrm{DLA}}$ has no parameters. Consider a set of observations of a QSO $\mathcal{D} = (\boldsymbol{\lambda}, \boldsymbol{y})$ with known observation noise variance vector $\boldsymbol{v}$. The model evidence for $\mathcal{M}_{\neg\mathrm{DLA}}$ given by observations can be computed directly:

$$p(\mathcal{D} \mid \mathcal{M}_{\neg\mathrm{DLA}}, \boldsymbol{v}, z_{\mathrm{QSO}}) \propto p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{v}, z_{\mathrm{QSO}}, \mathcal{M}_{\neg\mathrm{DLA}}). \tag{31}$$

The constant of proportionality is $p(\boldsymbol{\lambda} \mid \mathcal{M}_{\neg\mathrm{DLA}})$, a quantity that we do not model here. Rather, we will assume that $p(\boldsymbol{\lambda} \mid \mathcal{M})$ is constant across models, causing it to cancel during the calculation of the model posterior. Therefore, for the purposes of model comparison, we need only to compute

$$p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{v}, z_{\mathrm{QSO}}, \mathcal{M}_{\neg\mathrm{DLA}}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}, \mathbf{K} + \boldsymbol{\Omega} + \mathbf{V}), \tag{32}$$

where the $\boldsymbol{\mu}$, $\mathbf{K}$ and $\boldsymbol{\omega}$ learned above have been interpolated on to $\boldsymbol{\lambda}$.

6 A GP MODEL FOR QSO SPECTRAL SIGHTLINES WITH INTERVENING DLAS

In the previous section, we learned an appropriate GP model for QSO spectra without intervening DLAs, forming our null model $\mathcal{M}_{\neg\mathrm{DLA}}$. Here, we will extend that model to create a model for sightlines containing intervening DLAs. We will first fully describe the model for spectra containing exactly one intervening DLA, then extend this model to the case of two-or-more DLAs along a line of sight. We will call our model for lines of sight containing exactly k intervening DLAs $\mathcal{M}_{\mathrm{DLA(k)}}$; here we describe $\mathcal{M}_{\mathrm{DLA(1)}}$. Taking the conjunction of these models $\{\mathcal{M}_{\mathrm{DLA(i)}}\}_{i=1}^{\infty}$ gives our complete DLA model $\mathcal{M}_{\mathrm{DLA}}$.

Consider a quasar with redshift z_{QSO}, and suppose that there is an intervening DLA along the line of sight with redshift z_{DLA} and column density $N_{\mathrm{H\,I}}$. The effect of this on our observations is to multiply the emitted flux $f(\lambda)$ by an appropriate absorption function:

$$y(\lambda) = f(\lambda) \exp(-\tau(\lambda; z_{\mathrm{DLA}}, N_{\mathrm{H\,I}})) + \varepsilon, \tag{33}$$

where ε is additive Gaussian noise due to measurement error and τ is the absorption cross-section, which has a contribution corresponding to each transition we wish to model. Here, we model absorption for several members of the Lyman series:

$$\tau(\lambda; z_{\mathrm{DLA}}, N_{\mathrm{H\,I}}) = N_{\mathrm{H\,I}} \frac{\pi e^2 f \lambda'}{m_e c} \phi(v, b, \gamma), \tag{34}$$

where e is the elementary charge, λ' is the transition wavelength ($\lambda' = 1215.6701$ Å for Ly α) and f is the oscillator strength of the transition ($f = 0.4164$ for Ly α). The line profile function ϕ is a Voigt profile, where v is the relative velocity:

$$v = c\left(\frac{\lambda}{\lambda'(1 + z_{\mathrm{DLA}})} - 1\right), \tag{35}$$

$b/\sqrt{2}$ is the standard deviation of the Gaussian (Maxwellian) broadening contribution:

$$b = \sqrt{2\frac{kT}{m_p}}, \tag{36}$$

and γ is the width of the Lorenztian broadening contribution:

$$\gamma = \frac{\Gamma \lambda'}{4\pi}, \tag{37}$$

where Γ is a damping constant ($\Gamma = 6.265 \times 10^8\ \mathrm{s}^{-1}$ for Ly α). The gas temperature T is fixed to 10^4 K. This imparts a thermal broadening of 13 km s^{-1}, which is negligible compared to broadening of the DLA profile from Lorentzian damping wings. We neglect broadening due to any turbulence of the gas within the DLA, which could potentially contribute at lower column densities. We considered line profiles corresponding to Ly α, β and γ absorption, which we may compute for a given set of wavelengths given the known transition parameters, the temperature T, and z_{DLA} and $N_{\mathrm{H\,I}}$.

GPs provide a simple mechanism to model the multiplicative effect introduced by the absorption function $\exp(-\tau)$. Let a function f have a GP prior distribution $p(f) = \mathcal{GP}(f; \mu, K)$, and let $a(\lambda)$ be a known function. Then the distribution of the product

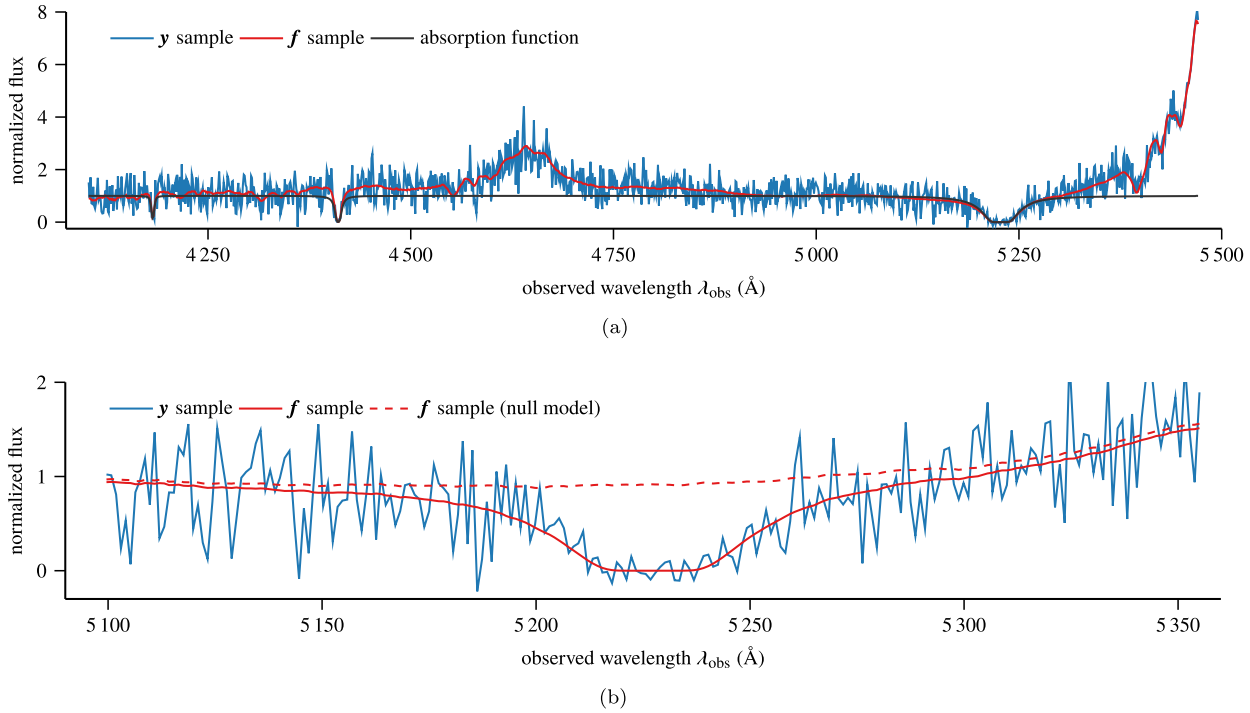


Figure 7. An example sample from our model for QSO emission spectra with one DLA along the line of sight. Here, we simulate a QSO with $z_{\rm QSO} = 2.5$ with a DLA at $z_{\rm DLA} = 2.2$ and $\log_{10} N_{\rm H\,I} = 20.8$. This sample corresponds to that in Fig. 6, but is instead drawn from the DLA model with the appropriate absorption profile (plotted in grey). In (a), we show the entire simulated observations, and in (b) we show detail in the region of the Ly α absorption central wavelength, with the continuum sample from Fig. 6 for comparison. Note that the full sample also reflects corresponding Ly β and Ly γ absorption.

$g(\lambda) = a(\lambda)f(\lambda)$ is also a Gaussian process (GPs are closed under affine transformations):

$$p(g) = \mathcal{GP}(f; \mu', K'), \tag{38}$$

where

$$\mu'(\lambda) = a(\lambda)\mu(\lambda); \qquad K'(\lambda, \lambda') = a(\lambda)K(\lambda, \lambda')a(\lambda'). \tag{39}$$

Therefore, given the parameters $(z_{\rm DLA}, N_{\rm H\,I})$ of a putative DLA, we compute the appropriate absorption function $\exp(-\tau(\lambda; z_{\rm DLA}, N_{\rm H\,I}))$ and modify the null GP model from the previous section as above. Specifically, consider observations of a QSO sightline at rest wavelengths $\boldsymbol{\lambda}$. Our model for the corresponding emitted flux $\boldsymbol{f}$ remains as in equation (11). Given the observation noise variance vector $\boldsymbol{v}$, the prior for the observation vector $\boldsymbol{y}$ without intervening DLAs is

$$p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{v}, z_{\rm QSO}, \mathcal{M}_{\neg \rm DLA}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}, \mathbf{K} + \boldsymbol{\Omega} + \mathbf{V}). \tag{40}$$

Suppose now that we wish to model the observed flux with a DLA at known redshift $z_{\rm DLA}$ and column-density $N_{\rm H\,I}$. First, we compute the theoretical absorption function with these parameters at $\boldsymbol{\lambda}$; call this vector $\boldsymbol{a}$:

$$\boldsymbol{a} = \exp(-\tau(\boldsymbol{\lambda}; z_{\rm DLA}, N_{\rm H\,I})). \tag{41}$$

Now, applying the result above, the prior for $\boldsymbol{y}$ with the specified DLA is

$$p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{v}, z_{\rm QSO}, z_{\rm DLA}, N_{\rm H\,I}, \mathcal{M}_{\rm DLA(1)}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{a} \circ \boldsymbol{\mu}, \mathbf{A}(\mathbf{K} + \boldsymbol{\Omega})\mathbf{A} + \mathbf{V}), \tag{42}$$

where $\boldsymbol{a} = \mathrm{diag}\mathbf{A}$.

Fig. 7 displays a draw from our DLA prior corresponding to the null model sample in Fig. 6.

An important feature of this model is that it is not in any way specific to DLAs, nor to data from the BOSS instrument. Our GP model for quasar emission spectra could be modified in an identical manner to model observed flux associated with any desired absorption feature.

6.1 Model evidence

Unlike our null model, which was parameter free, our DLA model $\mathcal{M}_{\rm DLA(1)}$ contains two parameters describing a putative DLA: the redshift $z_{\rm DLA}$ and column density $N_{\rm H\,I}$. We will denote the model parameter vector by $\theta = (z_{\rm DLA}, N_{\rm H\,I})$. To compute the model evidence, we must compute the following integral:

$$p(\mathcal{D} \mid \mathcal{M}_{\rm DLA(1)}, \boldsymbol{v}, z_{\rm QSO}) \propto p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{v}, z_{\rm QSO}, \mathcal{M}_{\rm DLA(1)}) = \int p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{v}, z_{\rm QSO}, \theta, \mathcal{M}_{\rm DLA(1)}) p(\theta \mid z_{\rm QSO}, \mathcal{M}_{\rm DLA(1)})\, \mathrm{d}\theta, \tag{43}$$

where we have marginalized the parameters given a prior distribution $p(\theta \mid z_{\rm QSO}, \mathcal{M}_{\rm DLA(1)})$. Before we describe the approximation of this integral, we will first describe the prior distribution used in our experiments.

6.2 Parameter prior

First, we make the assumption that absorber redshift and column density are conditionally independent given $z_{\rm QSO}$ and that the column density is independent of the QSO redshift:

$$p(\theta \mid z_{\rm QSO}, \mathcal{M}_{\rm DLA(1)}) = p(z_{\rm DLA} \mid z_{\rm QSO}, \mathcal{M}_{\rm DLA(1)}) p(N_{\rm H\,I} \mid \mathcal{M}_{\rm DLA(1)}). \tag{44}$$

For the distribution $p(z_{\mathrm{DLA}} \mid z_{\mathrm{QSO}}, \mathcal{M}_{\mathrm{DLA}(1)})$, we define the following range of allowable z_{DLA}:

$$z_{\min} = \max \begin{cases} \dfrac{\lambda_{\mathrm{Ly}\,\infty}}{\lambda_{\mathrm{Ly}\,\alpha}}(1 + z_{\mathrm{QSO}}) - 1 + 3000 \text{ km s}^{-1}/c \\ \dfrac{\min \lambda_{\mathrm{obs}}}{\lambda_{\mathrm{Ly}\,\alpha}} - 1 \end{cases} \tag{45}$$

$$z_{\max} = z_{\mathrm{QSO}} - 3000 \text{ km s}^{-1}/c; \tag{46}$$

that is, we insist the absorber centre be within the range of observed wavelengths (after restricting to $\lambda_{\mathrm{rest}} \in [911.75 \text{ Å}, 1216.75 \text{ Å}]$). We also apply a conservative cut-off of 3000 km s^{-1} in the immediate vicinity of the QSO to avoid proximity ionization effects, and in the immediate vicinity of the Lyman limit in the quasar rest frame (if visible) to avoid problems caused by possible incorrect determination of z_{QSO}.

Given these, we simply take a uniform prior distribution on this range:

$$p(z_{\mathrm{DLA}} \mid z_{\mathrm{QSO}}, \mathcal{M}_{\mathrm{DLA}(1)}) = \mathcal{U}[z_{\min}, z_{\max}]. \tag{47}$$

The column density prior $p(N_{\mathrm{H\,I}} \mid \mathcal{M}_{\mathrm{DLA}})$ is slightly more complicated. We first make a nonparametric estimate of the density, given the examples contained in the DLA catalogue provided with the BOSS DR9 Ly α forest sample. Due to the large dynamic range of column densities, we instead choose a prior on its base-10 logarithm, $\log_{10} N_{\mathrm{H\,I}}$.

We use the reported $\log_{10} N_{\mathrm{H\,I}}$ values for the $N_{\mathrm{DLA}} = 5854$ DLAs contained in the DR9 sample to make a kernel density estimate of the density $p(\log_{10} N_{\mathrm{H\,I}} \mid \mathcal{M}_{\mathrm{DLA}(1)})$. Kernel density estimation entails centring small so-called kernel functions on each observation and summing them to form our estimate. Here, we selected the univariate Gaussian probability density function for our kernels, with bandwidth selected via a plug-in estimator. The final estimate is

$$p_{\mathrm{KDE}}(\log_{10} N_{\mathrm{H\,I}} \mid \mathcal{M}_{\mathrm{DLA}(1)}) = \frac{1}{N_{\mathrm{DLA}}} \sum_{i=1}^{N_{\mathrm{DLA}}} \mathcal{N}(\log_{10} N_{\mathrm{H\,I}}; \ell_i, \sigma^2), \tag{48}$$

where ℓ_i is the base-10 logarithm of the ith observed column density. To account for some possible systematic bias in estimating this distribution, such as preferred numbers during visual inspection or underestimation of the probability of high-density systems due to low sample size, we make two adjustments. First, we simplify the form of the distribution by fitting a parametric prior to the nonparametric kernel density estimate of the form

$$\hat{p}_{\mathrm{KDE}}(\log_{10} N_{\mathrm{H\,I}} = N \mid \mathcal{M}_{\mathrm{DLA}(1)}) \quad q(\log_{10} N_{\mathrm{H\,I}} = N) \propto \exp(aN^2 + bN + c); \tag{49}$$

the values we learned, via least-squares fitting to the log probability over the range of $\log_{10} N_{\mathrm{H\,I}} \in [20, 22]$, were

$$a = -1.2695; \qquad b = 50.863; \qquad c = -509.33. \tag{50}$$

Finally, to account for some possible observation bias in the concordance catalogue, we take a mixture of this approximate column density prior with a simple log-uniform prior over a wide dynamic range:

$$p(\log_{10} N_{\mathrm{H\,I}} \mid \mathcal{M}_{\mathrm{DLA}(1)}) = \alpha q(\log_{10} N_{\mathrm{H\,I}} = N) + (1 - \alpha)\mathcal{U}[20, 23]. \tag{51}$$

Here, the mixture coefficient $\alpha = 0.9$ favours the data-driven prior. The upper limit of $\log_{10} N_{\mathrm{H\,I}} = 23$ is more than sufficient to model all thus-far observed DLAs. The final prior $p(\log_{10} N_{\mathrm{H\,I}} \mid \mathcal{M}_{\mathrm{DLA}(1)})$ is shown in Fig. 8), showing the expected bias towards smaller column densities.

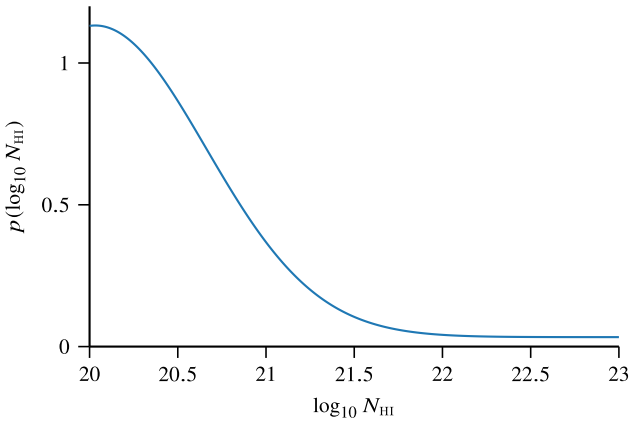


Figure 8. The probability density function of the log column density prior used in the experiments, $p(\log_{10} N_{\mathrm{H\,I}} \mid \mathcal{M}_{\mathrm{DLA}(1)})$.

6.3 Approximating the model evidence

Given our choice of parameter prior, the integral in equation (43) is unfortunately intractable, so we will result to numerical integration. In particular, we will use quasi-Monte Carlo (QMC) integration (Caflisch 1998). In QMC, we select N parameter samples $\{\theta_i\}$, evaluate the model likelihood given each of these samples, and approximate the integral in equation (43) by the sample mean:

$$p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{v}, z_{\mathrm{QSO}}, \mathcal{M}_{\mathrm{DLA}(1)}) \approx \frac{1}{N} \sum_{i=1}^{N} p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{v}, z_{\mathrm{QSO}}, \theta_i, \mathcal{M}_{\mathrm{DLA}(1)}). \tag{52}$$

This is the same estimator encountered in standard Monte Carlo integration, which selects the samples by sampling independently from the parameter prior $p(\theta \mid z_{\mathrm{QSO}}, \mathcal{M}_{\mathrm{DLA}(1)})$. QMC differs from normal Monte Carlo integration in that the samples $\{\theta_i\}$ are taken from a so-called *low-discrepancy sequence*, which guarantees the chosen samples are evenly distributed, leading to faster convergence. Here, we used $N = 10\,000$ samples generated from a scrambled Halton sequence (Kocis & Whiten 1997) to define our parameter samples. Note that the Halton sequence gives values approximately uniformly distributed on the unit square $[0, 1]^2$, which (after a trivial transformation) agrees in density with our redshift prior $p(z_{\mathrm{DLA}} \mid z_{\mathrm{QSO}}, \mathcal{M}_{\mathrm{DLA}(1)})$, but not our column density prior $p(\log_{10} N_{\mathrm{H\,I}} \mid \mathcal{M}_{\mathrm{DLA}(1)})$. To correct for this, we used inverse transform sampling to endow the generated samples with the appropriate distribution. For the inverse transformation, we used the approximated inverse cumulative distribution function corresponding to our prior in equation (51).

Note that we can use the same technique to approximate other quantities of interest. For example, if we wish to restrict our search to only DLAs with a certain minimum column density (e.g. $\log_{10} N_{\mathrm{H\,I}} > 22$), we can simply discard all parameter samples out of range, giving an unbiased estimate of the desired integral:

$$\int_{z_{\min}}^{z_{\max}} \int_{22}^{\infty} p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{v}, z_{\mathrm{QSO}}, \theta, \mathcal{M}_{\mathrm{DLA}(1)}) \times p(\theta \mid z_{\mathrm{QSO}}, \mathcal{M}_{\mathrm{DLA}(1)})\, \mathrm{d}z_{\mathrm{DLA}}\, \mathrm{d}\log_{10} N_{\mathrm{H\,I}}. \tag{53}$$

Note such estimators will, however, have higher variance due to the discarded parameter samples.

6.4 Multiple DLAS

While the catalogue we produce considers only one DLA per sightline, our model for QSO sightlines containing DLAs can readily model sightlines containing two or more intervening DLAs. Again, given the parameters ($z_{\rm DLA}$, $N_{\rm H\,I}$) of each absorber along the line of sight, we may compute the corresponding absorption function a and compute the observation posterior as in equation (42).

Let $\mathcal{M}_{\rm DLA(k)}$ represent a model explaining exactly k DLAs along the line of sight; we described $\mathcal{M}_{\rm DLA(1)}$ in the preceding sections. The model evidence integral (43) for $\mathcal{M}_{\rm DLA(k)}$ remains the same; however, θ will have dimension $2k$. Furthermore, we must consider the joint parameter prior $p(\theta \mid \mathcal{M}_{\rm DLA(k)})$.

We propose a (nearly) independent prior between each set of DLA parameters; the dependence will be discussed later. Rather than generating a $2k$-dimensional low-discrepancy sequence these parameters, we propose a stepwise approach. Given a spectrum, we first use the $\mathcal{M}_{\rm DLA(1)}$ parameter samples $\{\theta_i\}$ described above to approximate the model evidence (43). We can then approximate the posterior distribution of the single-DLA parameters by normalization:

$$p(\theta \mid z_{\rm QSO}, \mathcal{D}, \mathcal{M}_{\rm DLA(1)}) \propto p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{v}, z_{\rm QSO}, \theta_i, \mathcal{M}_{\rm DLA(1)}). \quad (54)$$

We may decompose the $\mathcal{M}_{\rm DLA(2)}$ parameters as $\theta = [\theta_1, \theta_2]^\top$, where each θ_i component describes a single DLA. We propose the following prior for the $\mathcal{M}_{\rm DLA(2)}$ model:

$$\begin{aligned}&(p(\theta_1, \theta_2 \mid z_{\rm QSO}, \mathcal{D}, \mathcal{M}_{\rm DLA(2)})\\ &\quad = p(\theta_1 \mid z_{\rm QSO}, \mathcal{D}, \mathcal{M}_{\rm DLA(1)}) p(\theta_2 \mid z_{\rm QSO}, \mathcal{M}_{\rm DLA(1)}).\end{aligned} \quad (55)$$

That is, we use the posterior probabilities from the analysis of the $\mathcal{M}_{\rm DLA(1)}$ model as the prior for the parameters of *one* of the DLAs when considering the $\mathcal{M}_{\rm DLA(2)}$ model. The prior for the parameters of the other DLA remains the noninformative prior as described above. For models $\mathcal{M}_{\rm DLA(k)}$ with $k > 2$, we apply a similar approach, where we combine a noninformative prior for θ_k with an informed prior for $\{\theta_i\}_{i=1}^{k-1}$:

$$\begin{aligned}&p\left(\{\theta\} \mid z_{\rm QSO}, \mathcal{D}, \mathcal{M}_{\rm DLA(k)}\right)\\ &\quad = p\left(\{\theta_i\}_{i=1}^{k-1} \mid z_{\rm QSO}, \mathcal{D}, \mathcal{M}_{\rm DLA(k-1)}\right) p(\theta_k \mid z_{\rm QSO}, \mathcal{M}_{\rm DLA(1)}).\end{aligned} \quad (56)$$

We do suggest injecting a small amount of dependence between the DLA parameters; specifically, any samples where any pair of $z_{\rm DLA}$ values correspond to a small relative velocity should be discarded to avoid samples describing two discrete DLAs in the same region of space.

In practice, the above scheme can be realized by first processing the spectrum with model $\mathcal{M}_{\rm DLA(1)}$; we then approximate the θ_1 posterior by renormalizing. To process the spectrum with model $\mathcal{M}_{\rm DLA(2)}$, we loop through the generated samples, each providing θ_2. For each sample, we sample a corresponding θ_1 sample from the approximate posterior. If the $z_{\rm DLA}$ values are too close, we discard the sample; otherwise, we have a valid θ sample with which to approximate the model evidence for $\mathcal{M}_{\rm DLA(2)}$. For $\mathcal{M}_{\rm DLA(k)}$, we proceed in a similar way, using some minor bookkeeping to approximate the $\{\theta_i\}_{i=1}^{k-1}$ posterior.

Note that the catalogue we produce considers only $\mathcal{M}_{\rm DLA(1)}$ to maintain statistical reliability with the low-SNR spectra from SDSS; however, the techniques we introduce are not tied to any particular source of data.

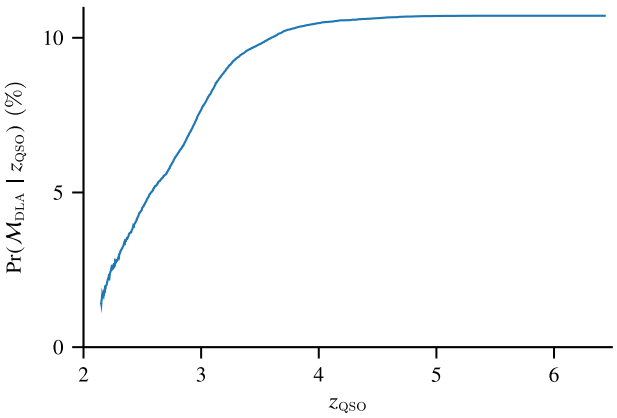


Figure 9. The redshift-dependent model prior $\Pr(\mathcal{M}_{\rm DLA} \mid z_{\rm QSO})$ computed from the BOSS DR9 Lyα forest sample with parameter $z' = 30\,000\,{\rm km\,s^{-1}}/c$.

7 MODEL PRIOR

Given a set of spectroscopic observations $\mathcal{D}$, our ultimate goal is to compute the probability the QSO sightline contains a DLA: $p(\mathcal{M}_{\rm DLA} \mid \mathcal{D})$. As described above, the Bayesian model selection approach requires two components: the data-independent prior probability that sightline contains a DLA, $\Pr(\mathcal{M}_{\rm DLA})$, and the ability to compute the ratio of model evidences $p(\mathcal{D} \mid \mathcal{M}_{\neg\rm DLA})$ and $p(\mathcal{D} \mid \mathcal{M}_{\rm DLA})$. The GP model built above allows us to compute the latter; in this section, we focus on the former.

Only approximately 10 per cent of the QSO sightlines in the DR9 release contain DLAs. A simple approach to prior specification would be to use a fixed value of $\Pr(\mathcal{M}_{\rm DLA}) \approx \frac{1}{10}$. However, it is less likely to observe a DLA in low-redshift QSOs due to the wavelength coverage of the SDSS and BOSS spectrographs being limited to $\lambda_{\rm obs} = 3800$ Å and $\lambda_{\rm obs} = 3650$ Å, respectively, on the blue end. Therefore, here we will use a slightly more sophisticated approach and derive a redshift-dependent prior $\Pr(\mathcal{M}_{\rm DLA} \mid z_{\rm QSO})$.

Our prior is simple and data driven. Consider a QSO with redshift $z_{\rm QSO}$. Let N be the number of QSOs in the training sample with redshift less than $z_{\rm QSO} + z'$, where z' is a small constant. Here, we took $z' = 30\,000\,{\rm km\,s^{-1}}/c$. Let M be the number of the sightlines of these containing DLAs within the range of quasar rest wavelengths we search here. We define

$$\Pr(\mathcal{M}_{\rm DLA} \mid z_{\rm QSO}) = \frac{M}{N}. \quad (57)$$

The constant z' serves to ensure that QSOs with very small redshift have sufficient data for estimating the prior. The resulting prior $\Pr(\mathcal{M}_{\rm DLA} \mid z_{\rm QSO})$ calculated from the DR9 sample is plotted in Fig. 9.

If we wish to break down our DLA prior $\Pr(\mathcal{M}_{\rm DLA} \mid z_{\rm QSO})$ into its component parts, for example to find $\Pr(\mathcal{M}_{\rm DLA(1)} \mid z_{\rm QSO})$, we assume that DLA occurrence is independent. If $\frac{M}{N}$ of sightlines contain at least one DLA, then $\frac{M^2}{N^2}$ contain at least two DLAs, etc., giving

$$\Pr(\mathcal{M}_{\rm DLA(k)} \mid z_{\rm QSO}) \approx \left(\frac{M}{N}\right)^k - \left(\frac{M}{N}\right)^{k+1}. \quad (58)$$

8 EXAMPLE

We have now developed all of the mathematical machinery required to compute the posterior odds that a given quasar sightline contains an intervening DLA, given a set of noisy spectroscopic observations $\mathcal{D}$. Briefly, we summarize the steps below, using the example from Fig. 1. We limit this example to searching for a single DLA, using only $\mathcal{M}_{\mathrm{DLA}(1)}$.

Consider a quasar with known redshift z_{QSO}, and suppose we have made spectroscopic observations of the object $\mathcal{D} = (\boldsymbol{\lambda}, \boldsymbol{y})$, with known observation noise variance vector $\boldsymbol{v}$. First, we compute the prior probability of the DLA model $\mathcal{M}_{\mathrm{DLA}}$, $\Pr(\mathcal{M}_{\mathrm{DLA}} \mid z_{\mathrm{QSO}})$ (57). This allows us to compute the prior odds in favour of the DLA model:

$$\frac{\Pr(\mathcal{M}_{\mathrm{DLA}} \mid z_{\mathrm{QSO}})}{\Pr(\mathcal{M}_{\neg\mathrm{DLA}} \mid z_{\mathrm{QSO}})} = \frac{\Pr(\mathcal{M}_{\mathrm{DLA}} \mid z_{\mathrm{QSO}})}{1 - \Pr(\mathcal{M}_{\mathrm{DLA}} \mid z_{\mathrm{QSO}})}. \tag{59}$$

For our example, $\Pr(\mathcal{M}_{\mathrm{DLA}} \mid z_{\mathrm{QSO}}) = 10.3\%$, giving prior odds of 0.114 (9-to-1 against the DLA model). Next, we compute the Bayes factor in favour of the DLA model:

$$\frac{p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{v}, z_{\mathrm{QSO}}, \mathcal{M}_{\mathrm{DLA}(1)})}{p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{v}, z_{\mathrm{QSO}}, \mathcal{M}_{\neg\mathrm{DLA}})}. \tag{60}$$

See equation (32) for how to compute the model likelihood for the null model and equation (43) for our approximation to the DLA model likelihood. For our DLA example, the Bayes factor overwhelmingly supports the DLA model, with a value of exp (96) $\approx 5 \times 10^{41}$. The computation of the Bayes factor is illustrated in Fig. 10, which shows the prior GP mean for the null model (Fig. 10a), the log likelihoods for the DLA model parameter samples (Fig. 10b), and the prior GP mean for the best DLA model parameter sample (Fig. 10c).

Finally, the posterior odds in favour of the sightline containing an intervening DLA is the product of equations (59) and (60). In practice, due to the typically large dynamic range of these quantities, it is numerically more convenient to compute the log odds. The log odds in favour of $\mathcal{M}_{\mathrm{DLA}}$ for the example from Fig. 1 are 94 nats,[9] and the probability of the sightline containing a DLA is effectively unity. The DLA parameter sample with the highest likelihood was $(z_{\mathrm{DLA}}, \log_{10} N_{\mathrm{H\,I}}) = (3.285, 20.33)$, closely matching the values reported in the DLA concordance catalogue $(z_{\mathrm{DLA}}, \log_{10} N_{\mathrm{H\,I}}) = (3.283, 20.39)$.

We may also compute the evidence for higher order models to derive a posterior distribution over the number of DLAs. In this case, the log model evidences for models $\mathcal{M}_{\mathrm{DLA}(2)}$, $\mathcal{M}_{\mathrm{DLA}(3)}$, $\mathcal{M}_{\mathrm{DLA}(4)}$ and $\mathcal{M}_{\mathrm{DLA}(5)}$ are -840, -977, -1141 and -1385, respectively, incorporating the model prior (57) and normalizing the single-DLA model dominates.

9 CATALOGUE

To verify the validity of our proposed method, we computed the posterior probability of $\mathcal{M}_{\mathrm{DLA}}$ for 162 858 quasar sightlines in the DR12Q release of SDSS III. Our catalogue and data products will be made available publicly at http://tiny.cc/dla_catalog_gp_dr12q, and the code to reproduce the entire catalogue from raw SDSS spectra will be posted under a permissive license at https://github.com/rmgarnett/gp_dla_detection.

[9] Nats are the logarithmic unit analogous to bits or dex corresponding to the base of the natural logarithm.

The full DR12Q catalogue contains 297 301 quasars, to which we applied the following cuts:

(i) We eliminate low-redshift ($z_{\mathrm{QSO}} < 2.15$) quasars. A total of 113 030 quasars in DR12Q satisfy this removal condition.

(ii) We eliminate broad absorption line (BAL) quasars, determined by the BAL visual inspection survey results in the `BAL_VI` field of the catalogue. A total of 29 580 quasars in DR12Q satisfy this removal condition.

(iii) We eliminate quasars that we cannot normalize due to no nonmasked pixels in the range of $\lambda_{\mathrm{rest}} \in [1310, 1325]$ Å. A total of 125 quasars in DR12Q satisfy this removal condition.

(iv) We eliminate quasars that have fewer than 200 nonmasked pixels in the range of $\lambda_{\mathrm{rest}} \in [911.75, 1216.75]$ Å. A total of 35 quasars in DR12Q satisfy this removal condition.

For each of the remaining spectra, we computed the posterior probability of the $\mathcal{M}_{\neg\mathrm{DLA}}$ and $\mathcal{M}_{\mathrm{DLA}(1)}$ models, given the observations, as described in the previous sections. We produce a full catalogue of our results, comprising two tables, the first rows of which are shown in Tables 1 and 2. The full catalogue will be available electronically alongside this manuscript.

When computing the likelihoods for the DLA model, we convolved the computed Voigt profile corresponding to each parameter sample with a Gaussian broadening profile with FWHM $= c/2000 = 150\,\mathrm{km\,s^{-1}}$, corresponding to the BOSS instrument's spectral resolution of $R \approx 2\,000$.

For each analysed spectrum , the result catalogue includes the following:

(i) The range of redshifts searched for DLAs, $[z_{\mathrm{min}}, z_{\mathrm{max}}]$

(ii) The log model prior, $\log \Pr(\mathcal{M} \mid z_{\mathrm{QSO}})$, for each model considered

(iii) The log model evidence, $\log p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{v}, z_{\mathrm{QSO}}, \mathcal{M})$, for each model considered

(iv) The model posterior, $\Pr(\mathcal{M} \mid \mathcal{D}, z_{\mathrm{QSO}})$

(v) The MAP estimates of the $\mathcal{M}_{\mathrm{DLA}(1)}$ model's parameters.

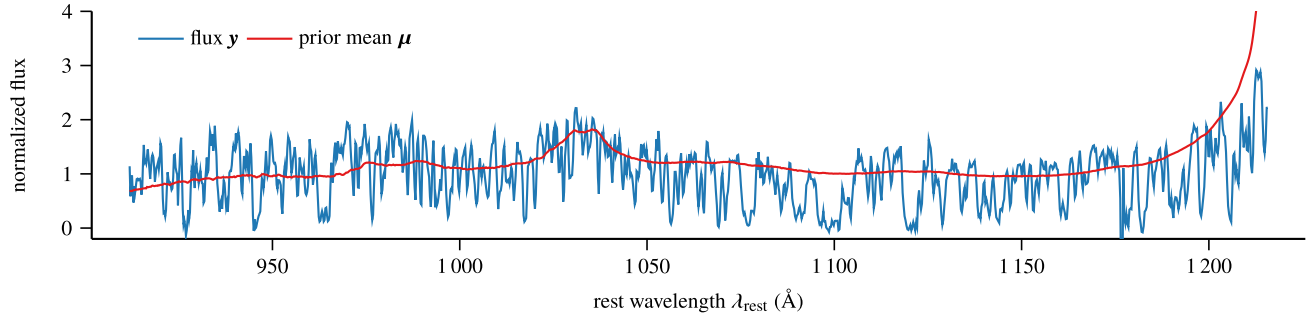
9.1 Running time

The running time of our approach allows it to easily scale to extremely large surveys and/or larger sample sizes. Our implementation is able to compute the model posterior over $\mathcal{M}_{\neg\mathrm{DLA}}$ and $\mathcal{M}_{\mathrm{DLA}(1)}$ in 0.5–2 seconds per spectrum on a standard Apple iMac desktop machine. For each spectrum, we must compute 10 001 log likelihoods of the form (32) [one for (32) and 10 000 for the $\mathcal{M}_{\mathrm{DLA}(1)}$ model (43)]; however, the low-rank structure of our covariance allows us to compute each rapidly using the identities in equations (22) and (23).

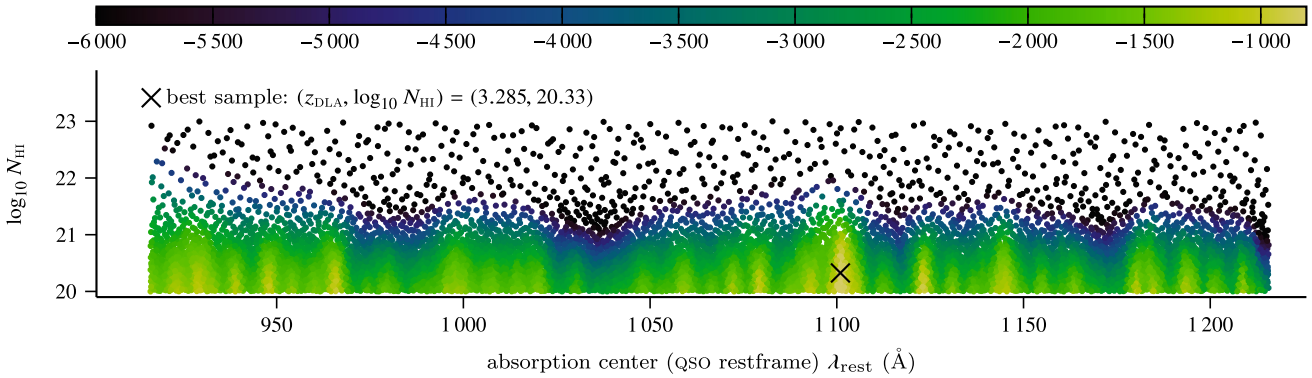
9.2 Analysis of results

To evaluate our results, we examined the ranking induced on the sightlines by the log posterior odds in favour of the DLA model $\mathcal{M}_{\mathrm{DLA}}$. If our method is performing correctly, true DLAs should be at the top of this list, above the non-DLA-containing sightlines. To visualize the quality of our ranking, we created a receiver-operating characteristic (ROC) plot, which, for every possible threshold on the log posterior odds, plots the false-positive rate (portion of non-DLAs with larger posterior odds) against the true positive rate (portion of DLAs with larger posterior odds).

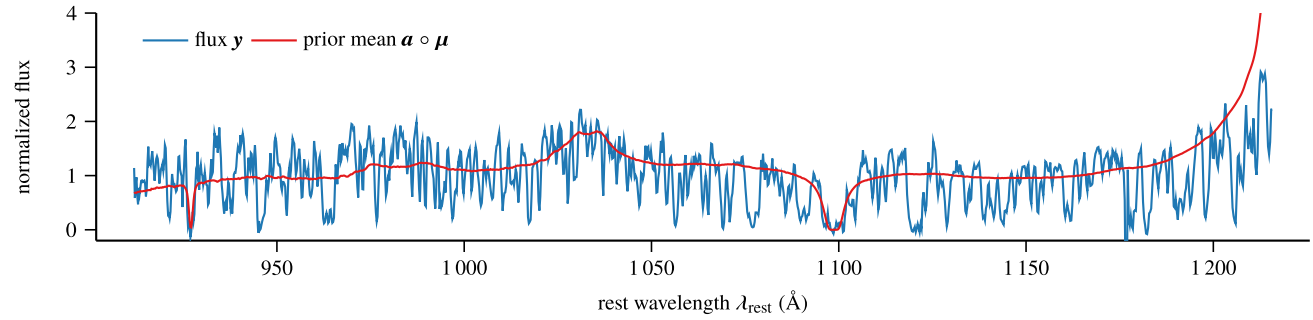
Notice that creating an ROC plot requires knowledge of the ground-truth labels for each of our objects, which of course we



(a) Null model: $\log p(\mathbf{y} \mid \boldsymbol{\lambda}, \mathbf{v}, z_{\text{QSO}}, \mathcal{M}_{\neg\text{DLA}}) = -908$



(b) Sample log likelihoods for DLA model, $\{\log p(\mathbf{y} \mid \boldsymbol{\lambda}, \mathbf{v}, z_{\text{QSO}}, \theta_i, \mathcal{M}_{\text{DLA}})\}$.



(c) DLA model: $\log p(\mathbf{y} \mid \boldsymbol{\lambda}, \mathbf{v}, z_{\text{QSO}}, \mathcal{M}_{\text{DLA}}) = -812$.

Figure 10. An illustration of the proposed DLA-finding procedure for the quasar sightline in Fig. 1. Panel (a) shows the normalized flux with the prior GP mean for our learned null model $\mathcal{M}_{\neg\text{DLA}}$. Panel (b) shows the log likelihoods for each of the parameter samples used to approximate the marginal likelihood of our DLA model $\mathcal{M}_{\text{DLA}}$. Panel (c) shows the normalized flux with the prior GP mean associated with the best DLA sample, $(z_{\text{DLA}}, \log_{10} N_{\text{H I}}) = (3.285, 20.33)$. Notice the Ly β absorption feature corresponding to this sample.

Table 1. The 297 301 objects in the SDSS-III DR12Q catalogue and the results of our cuts.

Thing id	SDSS name	Plate	MJD	fibre ID	Right ascension	Declination	z_{QSO}	SNR	Cut flags
268514930	000000.45+174625.4	6173	56238	0528	0.0018983	+17.7737391	2.3091	0.7795	0000

(297 300 rows removed)

do not have. Instead, we use the DLA concordance catalogue distributed with the BOSS DR9 Ly α forest catalogue as surrogate ground truth and restrict our analysis to lines of sight that both appear in that catalogue and were not removed by our cuts. A total of 54 360 objects comprise this intersection (99.9 per cent of the catalogue). The resulting ROC plot is displayed in Fig. 11. The top 1 per cent, 2 per cent, 5 per cent, 10 per cent and 20 per cent of our ranked list, respectively, recover 42.7 per cent, 57.5 per cent, 77.0 per cent, 89.1 per cent and 96.8 per cent of the DLAs listed in the concordance catalogue. Thus, even presorting the list by the posterior probability of $\mathcal{M}_{\text{DLA}}$ can dramatically speed up visual inspection.

A useful summary of the ROC plot is the area under the curve (AUC) statistic. The AUC has a natural interpretation: if we select a positive example and a negative example uniformly at random from those available, the AUC is the probability that the positive example

Table 2. The 162 858 objects in the SDSS-III DR12Q catalogue processed by our proposed GP DLA detection method, and a summary of derived quantities of interest.

	Search range		Model prior		Model evidence	
Thing id	z_{min}	z_{max}	$\log \Pr(\mathcal{M}_{\neg\text{DLA}} \mid z_{\text{QSO}})$	$\log \Pr(\mathcal{M}_{\text{DLA}} \mid z_{\text{QSO}})$	$\log p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{v}, z_{\text{QSO}}, \mathcal{M}_{\neg\text{DLA}})$	$\log p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{v}, z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)})$
268514930	1.9654	2.2989	-0.03081	−3.49537	−1.04359e+03	−1.04256e+03

Model posterior		$\underset{\theta}{arg\,max}\, p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{v}, z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)})$	
$\Pr(\mathcal{M}_{\neg\text{DLA}} \mid \mathcal{D}, z_{\text{QSO}})$	$\Pr(\mathcal{M}_{\text{DLA}} \mid \mathcal{D}, z_{\text{QSO}})$	z_{DLA}	$\log_{10} N_{\text{H I}}$
9.19661e-001	8.03389e-002	2.2160	20.0077

(162 857 rows removed)

Note: The first nine columns match Table 1 for the included objects (those with all cut flags equal to zero).

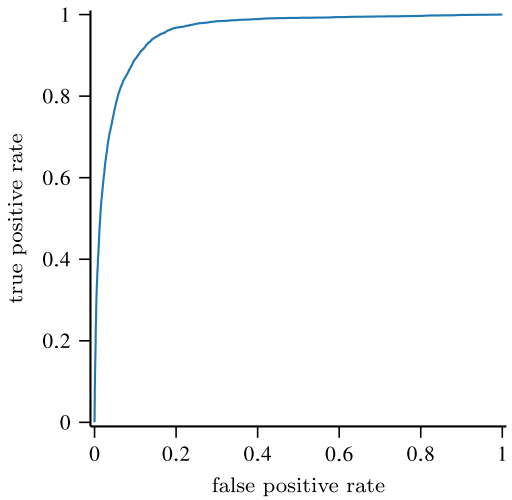


Figure 11. The ROC plot for the ranking of the 54 360 QSO sightlines contained in the BOSS DR9 Ly α forest sample (that were not filtered by our cuts), induced by the log posterior odds of containing a DLA. Ground-truth labellings were derived from the corresponding DLA concordance catalogue.

would be ranked higher than the negative example. For the DR9 DLA concordance catalogue surrogate, our AUC was 95.8 per cent. Clearly our approach is effective at identifying DLAs.

An important caveat to all of the results above is that none of the surrogates is likely to represent the true ground truth, and many 'false positive' sightlines could in fact contain as yet undiscovered DLAs. Fig. 12 gives an example of such a 'false positive,' showing the spectrum not contained in the DLA concordance catalogue that we rank the highest according to our model posterior ranking.

In fact, this spectrum appears to contain two DLAs along the line of sight. As a demonstration of our ability to detect multiple DLAs, we reprocessed this spectrum using the two-DLA model $\mathcal{M}_{\text{DLA}(2)}$. The data overwhelmingly support $\mathcal{M}_{\text{DLA}(2)}$ over either $\mathcal{M}_{\text{DLA}(1)}$ or $\mathcal{M}_{\neg\text{DLA}}$; $\Pr(\mathcal{M}_{\text{DLA}(2)} \mid \mathcal{D}, z_{\text{QSO}}) = 1 - 2.1 \times 10^{-22}$. Despite this line of sight not appearing in the DR9Q DLA concordance catalogue, we do note that it was flagged during the DR12Q visual inspection.

We have visually inspected several of these 'confident false positives'; of the top 30 such examples, 29 appear to contain large absorption features at the location indicated by the maximum likelihood parameter sample. The other is a very low-SNR spectrum that appears not to have been normalized satisfactorily.

The observation corresponding to our most egregious false negative, i.e. the spectrum flagged in the concordance catalogue that we assign the greatest confidence to being DLA free, is SDSS 081807.84+520935.1. There is a DLA along this line of sight, but outside the range of redshifts we search.

9.3 DLA parameter estimation analysis

The main goal of our DLA-detection method is to rank QSO sightlines by their probability of containing DLAs. The computation of the evidence of our DLA model $\mathcal{M}_{\text{DLA}}$ requires averaging over many samples of the DLA parameters $(z_{\text{DLA}}, \log_{10} N_{\text{H I}})$. We may use these samples to further derive point estimates of these parameters for presumed DLAs, if desired. The simplest approach is to report the sample with the highest likelihood:

$$\underset{\theta_i}{arg\,max}\, p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{v}, z_{\text{QSO}}, \theta_i, \mathcal{M}_{\text{DLA}}); \tag{61}$$

this represents the *maximum a posteriori* (MAP) estimate of the parameters. We analyse the behaviour of the MAP estimate by comparing it with the reported values in the DR9 concordance DLA catalogue.

The MAP estimates of the absorber redshift z_{DLA} are remarkably close to the catalogue figures. The median difference between the two is -2.7×10^{-4} ($-80.6\,\text{km}\,\text{s}^{-1}$) and the interquartile range is 2.5×10^{-3} ($742\,\text{km}\,\text{s}^{-1}$). Fig. 13(a) displays a kernel density estimate of the distribution of the difference between the MAP z_{DLA} estimates and the values reported in the concordance catalogue.

Examining the larger 'errors' in our estimation of z_{DLA}, we make an interesting observation that several z_{DLA} values reported in the concordance catalogue correspond exactly to the central wavelength of Ly β absorption for our redshift estimates. There does not seem to be an obvious pattern in the reverse direction, indicating that our method is less susceptible that previous techniques to mistaking Ly β absorption for Ly α absorption. Unlike previous approaches, which involve Voigt profile fitting to Ly α absorption only, we model the entire spectrum jointly, as well as the entire absorption profile corresponding to a given set of object parameters. Samples incorrectly setting z_{DLA} corresponding to a Ly β absorption feature should explain the observed spectrum worse than a sample setting z_{DLA} corresponding to a Ly α absorption feature, which in our set-up can better explain both the larger Ly α absorption and the corresponding Ly β feature.

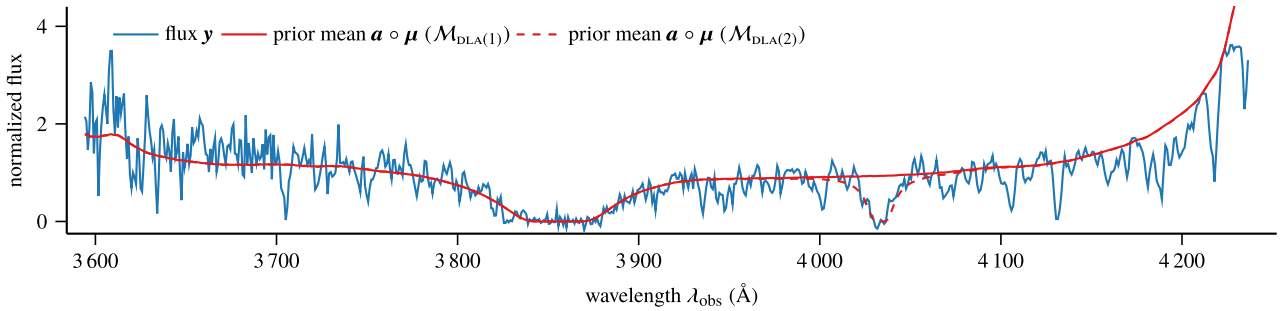


Figure 12. The spectrum appearing in the BOSS DR9 Ly α forest sample, not contained in the corresponding DLA concordance catalogue, with the highest posterior probability of containing an DLA according to our model. The object is SDSS 170023.94+205331.7, (plate, MJD, fibre) = (4175, 55680, 764), $z_{\mathrm{QSO}} = 2.4852$. We overwhelmingly believe there to be two DLAs along the line of sight; $\Pr(\mathcal{M}_{\mathrm{DLA}(2)} \mid \mathcal{D}, z_{\mathrm{QSO}}) = 1 - 2.1 \times 10^{-22}$. The prior means corresponding to the highest likelihood parameter sample for $\mathcal{M}_{\mathrm{DLA}(1)}$ and $\mathcal{M}_{\mathrm{DLA}(2)}$ are plotted, corresponding to $(z_{\mathrm{DLA}}, \log_{10} N_{\mathrm{H\,I}}) = (2.1717, 21.414)$ and $(z_{\mathrm{DLA}}, \log_{10} N_{\mathrm{H\,I}}) = \{(2.1715, 21.519), (2.3179, 20.075)\}$.

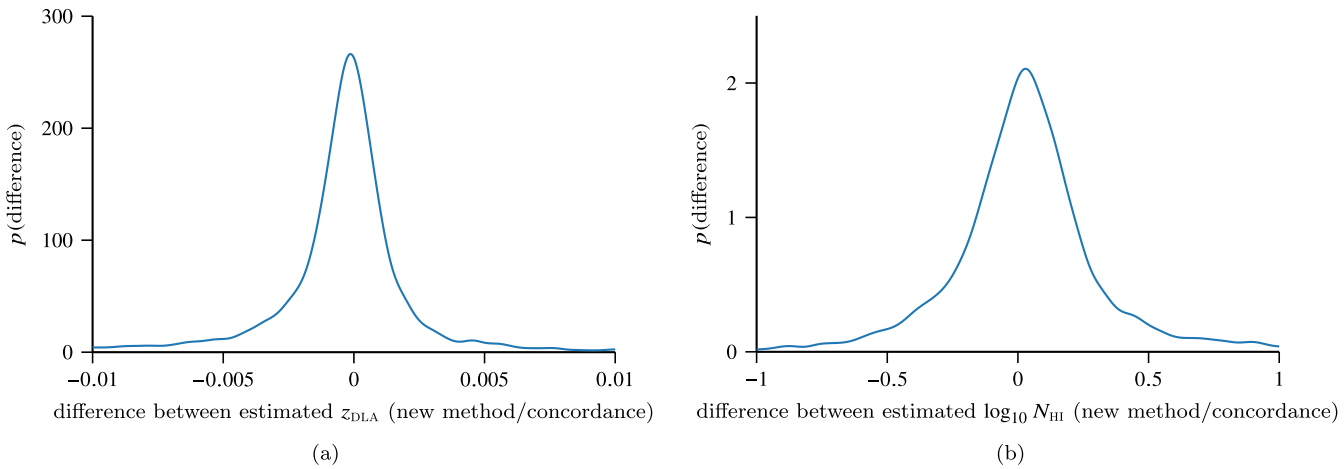


Figure 13. Kernel density estimate of the difference between the MAP estimates of the DLA parameters $(z_{\mathrm{DLA}}, \log_{10} N_{\mathrm{H\,I}})$ for DLAs listed in the BOSS DR9 Ly α forest sample, against the catalogue-reported values.

The MAP estimates of the log column density $\log_{10} N_{\mathrm{H\,I}}$ show more variation with the catalogue figures. The median difference between the two is quite small, only 0.030 dex. The interquartile range, however, is nontrivial at approximately 0.27 dex. Fig. 13(b) displays a kernel density estimate of the distribution of the difference between the MAP $\log_{10} N_{\mathrm{H\,I}}$ values and the values reported in the concordance catalogue.

In practice, for suspected DLAs, we suggest standard procedures for Voigt-profile fitting, if an accurate estimate of the parameters is desired. Our DLA-detection procedure is primarily concerned with the evidence contained in the entire set of parameter samples, and the MAP estimate carries no special significance. In particular, several parameter ranges might have large likelihood, corresponding to several potential absorption features. The MAP estimate alone cannot convey such information.

ACKNOWLEDGEMENTS

RG was supported by the National Science Foundation under Award Number IIA-1355406. SH and JS was supported by the US Department of Energy under Award Number DOE-DESC0011114. SB was supported by a McWilliams Fellowship from Carnegie Mellon University and by NASA through Einstein Postdoctoral Fellowship Award Number PF5-160133.

REFERENCES

Abazajian K. N. et al., 2009, ApJS, 182, 543
Ahn C. P. et al., 2012, ApJS, 203, 21
Ahn C. P. et al., 2014, ApJS, 211, 17
Anderson L. et al., 2012, MNRAS, 427, 3435
Anderson L. et al., 2014, MNRAS, 441, 24
Aubourg É. et al., 2015, Phys. Rev. D, 92, 123516
Bird S., Vogelsberger M., Haehnelt M., Sijacki D., Genel S., Torrey P., Springel V., Hernquist L., 2014, MNRAS, 445, 2313
Bird S., Haehnelt M., Neeleman M., Genel S., Vogelsberger M., Hernquist L., 2015, MNRAS, 447, 1834
Bird S., Garnett R., Ho S., 2017, MNRAS, 466, 2111
Bovy J. et al., 2011, ApJ, 729, 141
Caflisch R. E., 1998, Acta Numer., 7, 1
Carithers W., 2012, DLA Concordance Catalog, Published internally to SDSS
Cen R., 2012, ApJ, 748, 121
Chen H.-W., 2005, in Braun R., ed., ASP Conf. Ser. Vol. 331, Extra-Planar Gas. Astron. Soc. Pac., San Francisco, p. 371
Doi M. et al., 2010, AJ, 139, 1628
Eisenstein D. J. et al., 2011, AJ, 142, 72
Fukugita M., Ichikawa T., Gunn J. E., Doi M., Shimasaku K., Schneider D. P., 1996, AJ, 111, 1748
Fumagalli M., O'Meara J. M., Prochaska J. X., Rafelski M., Kanekar N., 2015, MNRAS, 446, 3178
Gardner J. P., Katz N., Weinberg D. H., Hernquist L., 1997, ApJ, 486, 42
Gunn J. E. et al., 1998, AJ, 116, 3040

Gunn J. E. et al., 2006, AJ, 131, 2332
Haehnelt M. G., Steinmetz M., Rauch M., 1998, ApJ, 495, 647
Jedamzik K., Prochaska J. X., 1998, MNRAS, 296, 430
Kocis L., Whiten W. J., 1997, ACM Transactions on Mathematical Software, 23, 266
Le Brun V., Bergeron J., Boisse P., Deharveng J. M., 1997, A&A, 321, 733
Lee K.-G. et al., 2013, AJ, 145, 69
Maller A. H., Prochaska J. X., Somerville R. S., Primack J. R., 2001, MNRAS, 326, 1475
Noterdaeme P. et al., 2012, A&A, 547, L1
Okoshi K., Nagashima M., 2005, ApJ, 623, 99
Pâris I. et al., 2012, A&A, 548, A66
Pâris I. et al., 2014, A&A, 563, A54
Pontzen A. et al., 2008, MNRAS, 390, 1349
Prochaska J. X., Wolfe A. M., 1997, ApJ, 487, 73
Prochaska J. X., Wolfe A. M., 2009, ApJ, 696, 1543
Prochaska J. X., Herbert-Fort S., Wolfe A. M., 2005, ApJ, 635, 123
Rahmati A., Pawlik A. H., Raicevic M., Schaye J., 2013, MNRAS, 430, 2427
Rao S. M., Nestor D. B., Turnshek D. A., Lane W. M., Monier E. M., Bergeron J., 2003, ApJ, 595, 94
Rasmussen C. E., Williams C. K. I., 2006, Gaussian Processes for Machine Learning. MIT Press, Cambridge, MA
Reid B. et al., 2016, MNRAS, 455, 1553
Ross N. P. et al., 2012, ApJS, 199, 3
Slosar A. et al., 2011, J. Cosmology Astropart. Phys., 9, 1
Smee S. A. et al., 2013, AJ, 146, 32
Smith J. A. et al., 2002, AJ, 123, 2121
Wolfe A. M., Turnshek D. A., Smith H. E., Cohen R. D., 1986, ApJS, 61, 249
Wolfe A. M., Gawiser E., Prochaska J. X., 2005, ARA&A, 43, 861
York D. G. et al., 2000, AJ, 120, 1579

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

Table 1. The 297 301 objects in the SDSS-III DR12Q catalogue, and the results of our cuts.
Table 2. The 162 858 objects in the SDSS-III DR12Q catalogue processed by our proposed GP DLA detection method, and a summary of derived quantities of interest.
table1.dat
table2.dat

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

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