Finding Galaxies in the Shadows of Quasars with Gaussian Processes

Roman Garnett GARNETT@WUSTL.EDU
Washington University in St. Louis, St. Louis, MO 63130, United States

Shirley Ho SHIRLEYH@ANDREW.CMU.EDU
Jeff Schneider JEFF.SCHNEIDER@CS.CMU.EDU
Carnegie Mellon University, Pittsburgh, PA 15213, United States

Abstract

We develop an automated technique for detecting damped Lyman-α absorbers (DLAs) along spectroscopic sightlines to quasi-stellar objects (QSOs or quasars). The detection of DLAs in large-scale spectroscopic surveys such as SDSS–III is critical to address outstanding cosmological questions, such as the nature of galaxy formation. We use nearly 50 000 QSO spectra to learn a tailored Gaussian process model for quasar emission spectra, which we apply to the DLA detection problem via Bayesian model selection. We demonstrate our method's effectiveness with a large-scale validation experiment on over 100 000 spectra, with excellent performance.

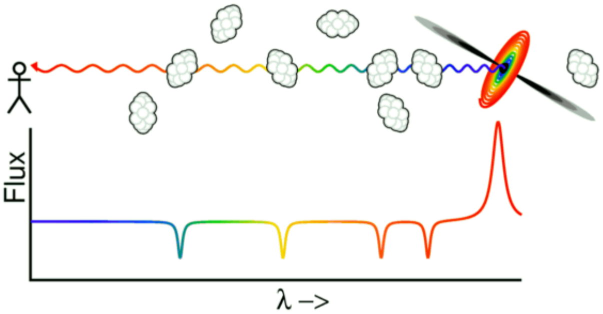


Figure 1. A cartoon of Lyman-α emission and absorption (Wright, 2004). Lyman-α photons are emitted by a high-redshift quasar (top right). On their way to Earth, these photons travel through clouds containing neutral hydrogen gas, each at a different redshift. Each may absorb some photons, tracing an absorption feature in the measured spectrum; these together form the Lyman-α forest.

1. Introduction

Damped Lyman-α systems (DLAs) are large gaseous objects containing large amounts of neutral hydrogen (HI) gas. DLAs emit little light and cannot be observed directly; however, they can be detected indirectly in the spectroscopic measurements of high-redshift quasi-stellar objects (QSOs or quasars), due to their leaving telltale wide absorption phenomena.

DLAs currently represent our only probe of normal (neither high-mass nor star-forming) galaxies at high redshift. DLAs 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 (Wolfe et al., 2005). These systems likely played a significant role in fueling star formation across the cosmic time. Due to their importance, it is common practice for astronomers to visually inspect every measured quasar spectrum to identify potential DLAs. This is a daunting task: the large-scale Sloan Digital Sky Survey III (SDSS–III) has measured nearly 300 000 quasar spectra over its brief history (Eisenstein et al., 2011). With several large-scale spectroscopic surveys of QSOs due to start soon (including SDSS–IV and DESI[1]), which plan to observe 1–2 million quasars, it would be unimaginable to continue the status quo of visually inspecting every spectrum. We present a fully automated and scalable method to find DLAs in quasar spectra, based on Gaussian processes and Bayesian model selection, using a massive dataset from the SDSS–III project.[2]

1.1. Relevant Spectroscopic Concepts

In spectroscopy, we use an instrument to measure the *spectral flux* (electromagnetic radiation emitted per unit area per unit wavelength) of an object over a range of wavelengths of light, binned into discrete intervals called "pixels." Due to the expansion of the Universe, flux corresponding to a wavelength λ_{rest} in the restframe of an observed object will be observed on Earth at a redshifted wavelength λ_{obs}; the relationship between these quantities is $\lambda_{\text{obs}} = (1 + z)\lambda_{\text{rest}}$, where z is the cosmological redshift.

Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

[1] http://desi.lbl.gov/
[2] http://www.sdss.org/

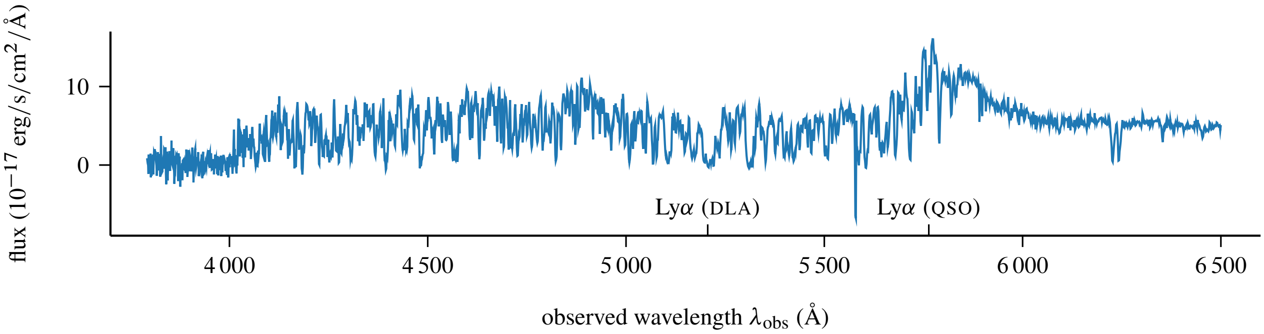


Figure 2. A portion of an example spectrum for the object SDSS 020712.80+052753.4, $z_{\text{QSO}} = 3.741$. The Lyα emission is marked, occurring at $\lambda_{\text{obs}} = 5\,763$ Å. This QSO is included in the DLA concordance catalog with $(z_{\text{DLA}}, \log_{10} N_{\text{HI}}) = (3.283, 20.39)$, corresponding to central absorption wavelength $\lambda_{\text{obs}} = 5\,206$ Å or $\lambda_{\text{rest}} = 1\,098$ Å in the QSO restframe.

Quasar emission spectra contain numerous *emission lines*, large localized spikes in flux, corresponding to well-understood atomic events. The most important of these in the context of DLAs is the *Lyman series,* a series of emission lines corresponding to photons emitted from a neutral hydrogen atom when its electron transitions from a higher-energy orbital $n > 1$ to the ground state $n = 1$. Due to quantum mechanics, each of these transitions corresponds to a specific amount of energy being released; this energy is conserved via the emission of a photon of predictable wavelength. The Lyman-α (Lyα) transition is the $n = 2 \to n = 1$ member of this series, which corresponds to a photon emission with wavelength $\lambda_{\text{rest}} = 1\,216$ Å. Lyman-β (Lyβ) is the $n = 3 \to n = 1$ transition, etc. The *Lyman limit* (Ly∞) at $\lambda_{\text{rest}} = 912$ Å is the highest-energy member of the series, equivalent to the energy required to strip the electron from the atom completely.

The line of sight from Earth to a quasar must pass through a vast expanse of the intergalactic medium. Along the line of sight will be numerous objects containing neutral hydrogen. When a Lyα photon emitted by the quasar passes through such an object, it can be absorbed by a neutral hydrogen atom, exciting it (inducing the reverse $n = 1 \to n = 2$ transition of its electron). Such absorption causes corresponding dips in the observed flux. Again, due to the expansion of the Universe, each of these objects will be located at a different redshift, necessarily less than the quasar itself. The result is the so-called *Lyman-α forest*, a series of dips in quasar emission spectra bluewards (that is, at lower wavelengths) from the Lyα emission line. Figure 1 shows a cartoon illustration of this process.

When a very dense cloud containing neutral hydrogen gas (column density surpassing $N_{\text{HI}} > 2 \times 10^{20}\,\text{cm}^{-2}$ along the line of sight), the absorption profile exhibits characteristic "damping wings," and the object is classified a damped Lyman-α absorber (DLA). Figure 2 shows a an example quasar spectrum gathered by the SDSS–III project that contains a DLA along the line of sight. The entire Lyman-α forest is also visible.

1.2. Notation

We will briefly establish some notation. Consider a QSO with redshift z_{QSO}; we will always assume that z_{QSO} is known, allowing us to work in the quasar restframe. 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 rather than observed wavelengths. Note that the 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 $y(\lambda)$.

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 an intervening DLA ($\mathcal{M}_{\text{DLA}}$), and one for those without ($\mathcal{M}_{\neg\text{DLA}}$). Then, given the available data, we will compute the posterior probability that the former model is correct. Both models will be based on Gaussian processes, which we describe below.

2. 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. *Gaussian processes* (GPs) provide a powerful nonparametric framework for modeling unknown functions, which we will adopt for this task. See (Rasmussen & Williams, 2006) for an extensive introduction to GPs.

Let $\mathcal{X}$ be an arbitrary input space, for example the real line $\mathbb{R}$, and let $f\colon \mathcal{X} \to \mathbb{R}$ be a real-valued function on $\mathcal{X}$ we wish to model. Given an arbitrary mean function $\mu\colon \mathcal{X} \to \mathbb{R}$ and positive semidefinite covariance function $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we may endow f with a Gaussian process prior distribution $p(f) = \mathcal{GP}(f; \mu, K)$. The defining characteristic of a GP is that given a finite set of inputs $\boldsymbol{\lambda}$, the corresponding vector of function values $\mathbf{f} = f(\boldsymbol{\lambda})$ is multivariate Gaussian distributed:

$$p(\mathbf{f}) = \mathcal{N}\big(\mathbf{f}; \mu(\boldsymbol{\lambda}), K(\boldsymbol{\lambda}, \boldsymbol{\lambda})\big). \tag{1}$$

Consider a set of noisy observations of f at $\boldsymbol{\lambda}$, $\mathcal{D} = (\boldsymbol{\lambda}, \mathbf{y})$. We will assume these observations are generated by corrupting the true latent values $\mathbf{f}$ by zero-mean, independent Gaussian noise. We assume that the noise variance associated with each of our measurements is known[3] and given by a corresponding vector $\boldsymbol{\nu}$, with $\nu_i = \sigma(\lambda_i)^2$. Given the noise independence assumption, the entire observation model is given by

$$p(\mathbf{y} \mid \boldsymbol{\lambda}, \mathbf{f}, \boldsymbol{\nu}) = \mathcal{N}(\mathbf{y}; \mathbf{f}, \mathbf{N}), \tag{2}$$

where $\mathbf{N} = \operatorname{diag} \boldsymbol{\nu}$. 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 emission spectra, where noise levels can vary widely as a function of observed wavelength.

Given our GP prior on f (1) and the Gaussian noise observation model (2), we may compute the *marginal likelihood* of the data in closed form:

$$p(\mathbf{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}) = \mathcal{N}\big(\mathbf{y}; \mu(\boldsymbol{\lambda}), K(\boldsymbol{\lambda}, \boldsymbol{\lambda}) + \mathbf{N}\big).$$

In typical applications of GP inference, the prior mean μ and prior covariance K would be selected from numerous off-the-shelf solutions available; however, none of these would be directly appropriate for modeling QSO emission spectra, due to their complex nature. For example, strong off-diagonal correlations must exist between potentially distant emission lines, such as members of the Lyman series. Rather, we will construct a custom GP prior distribution for modeling these spectra in the next section.

3. Learning a GP Model for QSO Spectra

We wish to construct a Gaussian process 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 $\mathcal{M}_{\neg\text{DLA}}$. We will later extend this to form our DLA model $\mathcal{M}_{\text{DLA}}$.

A Gaussian process is defined entirely by its first two moments, μ and K. Our goal in this section will be to derive reasonable choices for these functions. We adopt a data-driven approach and learn an appropriate model given over 48 000 examples contained in a previously compiled catalog of quasar spectra recorded by the BOSS spectrograph (Smee et al., 2013).

3.1. Description of Data

We used the QSO spectra from the BOSS DR9 Lyman-α forest sample (Lee et al., 2013) to train our GP model. Later, we will use the spectra from the corresponding DR10 sample (Lee et al., 2014) to evaluate our proposed DLA finding approach. The DR9 sample comprises 54 468 QSO spectra with $z_{\text{QSO}} \geq 2.15$ from the SDSS DR9 release appropriate for Lyman-α forest analysis; the DR10 sample comprises 101 167 spectra. 53 490 QSOs are contained in both catalogs.

Both catalogs have been augmented with a previously compiled "concordance" DLA catalog (Carithers, 2012), combining the results of three previous DLA searches. These include a visual-inspection survey (Slosar et al., 2011) and two previous semi-automated approaches: a template-matching approach (Noterdaeme et al., 2012), and an approach using Fisher discriminant analysis (Carithers, 2012). Any sightline flagged in at least two of these catalogs is included in the concordance. A total of 5 854 lines of sight are flagged as containing an intervening DLA in the DR9 catalog (10.7%); 9 531 are flagged in the DR10 catalog (9.4%).

3.2. Modeling 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 by dividing by the median flux observed bluewards of the Lyα emission.

We model emissions in the range $\lambda_{\text{rest}} \in [800\,\text{Å}, 1\,216\,\text{Å}]$, where DLAs are most likely to be observed. 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. The chosen grid was this set of wavelengths with a linearly equal spacing of $\Delta\lambda = 0.25\,\text{Å}$. This resulted in a vector of input locations $\boldsymbol{\lambda}$ with $|\boldsymbol{\lambda}| = N_{\text{pixels}} = 1\,665$ pixels.

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

$$p(\mathbf{f} \mid \boldsymbol{\lambda}, z_{\text{QSO}}) = \mathcal{N}(\mathbf{f}; \boldsymbol{\mu}, \mathbf{K}), \tag{3}$$

where $\boldsymbol{\mu} = \mu(\boldsymbol{\lambda})$ and $\mathbf{K} = K(\boldsymbol{\lambda}, \boldsymbol{\lambda})$. Note that we must

[3] This is a valid assumption for spectroscopic surveys such as SDSS–III, which typically generate pixelwise noise estimates for each measurement.

condition on the QSO redshift z_{QSO} because it is required for shifting into the quasar restframe.

As mentioned previously, however, we can never observe f directly, both due to measurement error and absorption by intervening matter. The former can be handled easily for our spectra by using the SDSS pipeline error estimates in the role of the noise vector $\boldsymbol{\nu}$. The latter is more problematic, especially in our selected region, which includes the Lyman-α forest. In principle, if we knew the exact nature of the intervening matter, we could model this absorption explicitly, but this is unrealistic. We will instead model the effect of small absorption phenomena (absorption by objects with column density below the DLA limit) by an additive wavelength-dependent Gaussian noise term, which we will learn. Therefore the characteristic "dips" of the Lyman-α forest will be modeled as noisy deviations from the true underlying smooth continuum. Later we will explicitly model larger absorption phenomena to build our DLA model.

The mathematical consequence of this modeling decision is as follows. Consider the arbitrary GP model in (3). We wish to model the associated spectroscopic observation values on the chosen grid, $\mathbf{y} = y(\boldsymbol{\lambda})$ with measurement noise vector $\boldsymbol{\nu} = \sigma(\boldsymbol{\lambda})^2$. The model we adopt here will involve a global additional noise vector $\boldsymbol{\omega}$ modeling absorption deviations from the quasar continuum. The observation model resulting from this choice is

$$p(\mathbf{y} \mid \mathbf{f}, \boldsymbol{\nu}, \boldsymbol{\omega}, \mathcal{M}_{\neg\text{DLA}}) = \mathcal{N}(\mathbf{y}; \mathbf{f}, \boldsymbol{\Omega} + \mathbf{N}),$$

where $\boldsymbol{\Omega} = \operatorname{diag} \boldsymbol{\omega}$. Therefore, given our chosen input grid $\boldsymbol{\lambda}$, the prior distribution of associated spectroscopic observations $\mathbf{y}$ is

$$p(\mathbf{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, \boldsymbol{\omega}, z_{\text{QSO}}, \mathcal{M}_{\neg\text{DLA}}) = \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}, \mathbf{K}+\boldsymbol{\Omega}+\mathbf{N}). \quad (4)$$

Our goal now is to learn appropriate values for $\boldsymbol{\mu}$, $\mathbf{K}$, and $\boldsymbol{\omega}$, which will fully specify our null model $\mathcal{M}_{\neg\text{DLA}}$.

3.3. Learning Appropriate Parameters

To build our null model, we took the $N_{\text{spec}} = 48\,614$ spectra from the BOSS DR9 Lyman-α forest sample that are putatively absent of intervening DLAs. We linearly interpolated the normalized flux and noise variance measurements of each spectrum onto the chosen wavelength grid $\boldsymbol{\lambda}$, after masking problematic pixels as indicated by the SDSS pipeline. Note that we did not "interpolate through" masked pixels. We also did not extrapolate beyond the range of wavelengths present in each spectrum.

We collect the resulting vectors into $(N_{\text{spec}} \times N_{\text{pixels}})$ matrices $\mathbf{Y}$ and $\mathbf{N}$, containing the normalized flux and noise variance vectors, respectively. For QSO i, we will write $\mathbf{y}_i$ and $\boldsymbol{\nu}_i$ to represent the corresponding observed flux and noise variance vectors, and will define $\mathbf{N}_i = \operatorname{diag} \boldsymbol{\nu}_i$. Due to masked pixels and varying redshifts of each QSO, these matrices contain numerous missing values, especially on the blue end.

3.3.1. Learning the Mean

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

$$\mu_j = \underset{y_{ij} \neq \text{NaN}}{\text{median}}\, y_{ij}.$$

Note that the sample mean is the maximum-likelihood estimator for $\boldsymbol{\mu}$; however, here we used the median to be more robust to large outliers that typically appear on the blue end of QSO spectra. The learned mean vector $\boldsymbol{\mu}$ is plotted in Figure 3.

3.3.2. Learning the Covariance

We will use standard unconstrained optimization techniques to learn the covariance matrix $\mathbf{K}$ and absorption "noise" vector $\boldsymbol{\omega}$. We use a low-rank decomposition to limit the number of free variables in our model:

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

where $\mathbf{M}$ is an $(N_{\text{pixels}} \times k)$ matrix with $k \ll N_{\text{pixels}}$. This decomposition allows unconstrained optimization for $\mathbf{M}$. Here we took $k = 10$, noting that the first 10 principal components of the flux matrix $\mathbf{Y}$ explain approximately 99.9% of the total variance.

We assume that our measured flux vectors are independent realizations drawn from a common prior (4):

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

where all missing values are ignored. That is, in the ith entry of the product, we only use the entries of $\boldsymbol{\mu}$, $\boldsymbol{\nu}_i$, and $\boldsymbol{\omega}$, and only the rows of $\mathbf{M}$, corresponding to non-masked values in $\mathbf{y}_i$.

We define the log likelihood of the data, $\mathcal{L}$, as a function of the covariance parameters $\mathbf{M}$ and $\boldsymbol{\omega}$:

$$\mathcal{L}(\mathbf{M}, \boldsymbol{\omega}) = \log p(\mathbf{Y} \mid \boldsymbol{\lambda}, \boldsymbol{\mu}, \mathbf{M}, \boldsymbol{\omega}, \mathbf{N}, \mathbf{z}_{\text{QSO}}, \mathcal{M}_{\neg\text{DLA}}).$$

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 parameterize $\boldsymbol{\omega}$ entrywise by its natural logarithm.

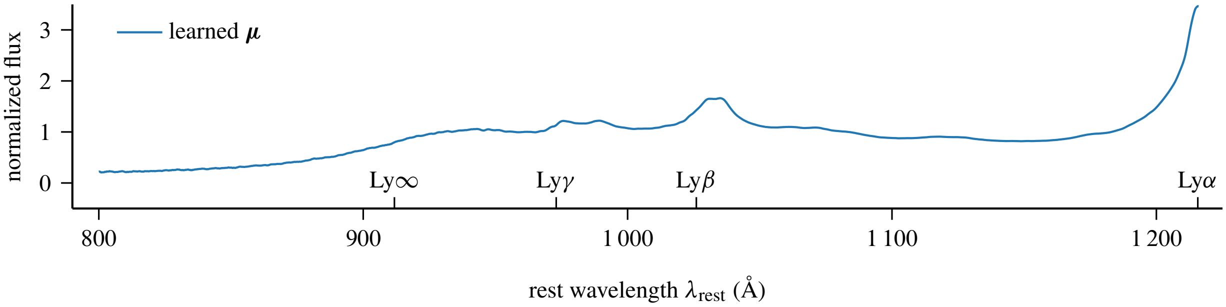


Figure 3. The learned mean vector μ derived by taking the median across the stacked spectra. The vector has been smoothed with a 11-pixel (2.75 Å) boxcar function for clarity on the blue end. Members of the Lyman series are marked.

An important feature of our particular choice of model is that we can solve linear systems with and compute the log determinant of the prior observation covariance

$$\mathbf{K} + \mathbf{\Omega} + \mathbf{N}_i$$

quickly via the Woodbury identity and the Sylvester determinant theorem, respectively. This allows very fast computation of $\mathcal{L}$ despite the very large number of training spectra and pixels.

To maximize our joint log likelihood, we applied the L-BFGS algorithm, a gradient-based quasi-Newton algorithm for unconstrained optimization. The initial value for $\mathbf{M}$ was taken to be the top-10 principal components of $\mathbf{Y}$; the initial value of each entry in $\boldsymbol{\omega}$ was taken to be the log sample variance of the corresponding column of $\mathbf{Y}$, ignoring missing values.

The learned prior covariance matrix $\mathbf{M}\mathbf{M}^\top + \mathbf{\Omega}$ is shown in Figure 4 (normalized to show correlations rather than raw covariances). Features corresponding to the Lyman series are clearly visible, including strong off-diagonal correlations between pairs of emission lines and an especially prominent correlation feature above the Lyman limit, Ly∞. Such features are highly atypical of off-the-shelf covariances.

To apply our model to spectroscopic 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 $\boldsymbol{\omega}$ onto the desired wavelengths. We may also account for redshift trivially should we wish to work in the λ_{obs} domain.

3.4. Model Evidence

We note that our null model $\mathcal{M}_{\neg\text{DLA}}$ has no parameters beyond those already learned and fixed. Consider a set of spectroscopic observations of a QSO sightline $\mathcal{D} = (\boldsymbol{\lambda}, \mathbf{y})$ with known observation noise variance vector $\boldsymbol{\nu}$. The model evidence for $\mathcal{M}_{\neg\text{DLA}}$ given by these observations can be computed directly:

$$p(\mathbf{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}, \mathcal{M}_{\neg\text{DLA}}) = \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}, \mathbf{K} + \mathbf{\Omega} + \mathbf{N}), \quad (5)$$

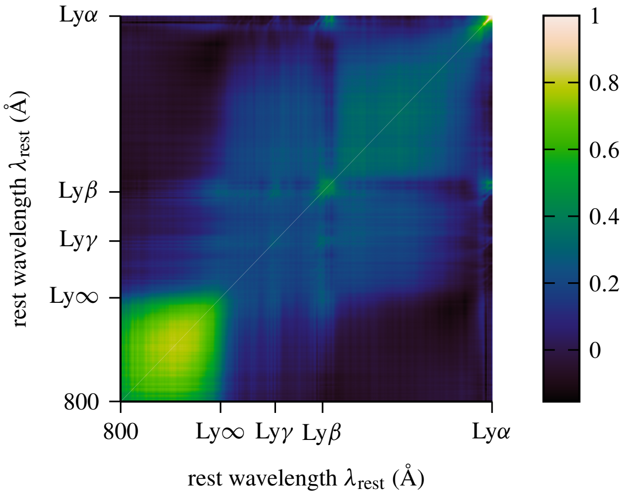


Figure 4. The observation correlation matrix (normalized $\mathbf{K} + \mathbf{\Omega}$) corresponding to the learned parameters.

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

4. A Model for Intervening DLAs

In the previous section, we learned a GP model for QSO spectra without intervening DLAs. Here we will extend that model to create a model for sightlines containing intervening DLAs.

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

$$y(\lambda) = f(\lambda)\exp\bigl(-\tau(\lambda; z_{\text{DLA}}, N_{\text{HI}})\bigr) + \varepsilon(\lambda),$$

where $\varepsilon(\lambda)$ is additive Gaussian noise due to measurement error and τ is a *Voigt profile* for Lyman-α absorption for the specified system, which can be computed explicitly via physical modeling (Draine, 2011).

Thankfully, Gaussian processes are closed under linear transformation. Suppose that we wish to model the observed flux along the sightline with a DLA with known redshift and column density. First we compute the theoretical absorption function with these parameters at $\boldsymbol{\lambda}$; call this vector $\mathbf{a}$:

$$\mathbf{a} = \exp\bigl(-\tau(\boldsymbol{\lambda}; z_{\text{DLA}}, N_{\text{HI}})\bigr).$$

Now the prior for $\mathbf{y}$ with the specified DLA is

$$p(\mathbf{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}, z_{\text{DLA}}, N_{\text{HI}}, \mathcal{M}_{\text{DLA}}) \\ = \mathcal{N}\bigl(\mathbf{y}; \mathbf{a} \circ \boldsymbol{\mu}, \mathbf{A}(\mathbf{K} + \boldsymbol{\Omega})\mathbf{A} + \mathbf{N}\bigr),$$

where $\mathbf{a} = \operatorname{diag} \mathbf{A}$ and $\circ$ represents the Hadamard (elementwise) product.

4.1. Model Evidence

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

$$p(\mathbf{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}, \mathcal{M}_{\text{DLA}}) = \\ \int p(\mathbf{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}, \theta, \mathcal{M}_{\text{DLA}}) p(\theta \mid z_{\text{QSO}}, \mathcal{M}_{\text{DLA}}) \, \mathrm{d}\theta, \quad (6)$$

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

4.2. Parameter Prior

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

$$p(\theta \mid z_{\text{QSO}}, \mathcal{M}_{\text{DLA}}) = p(z_{\text{DLA}} \mid z_{\text{QSO}}, \mathcal{M}_{\text{DLA}}) p(N_{\text{HI}} \mid \mathcal{M}_{\text{DLA}}).$$

We define the following range of allowable z_{DLA}:

$$z_{\min} = (\min \lambda_{\text{obs}})/(1\,216\,\text{Å}) - 1; \qquad z_{\max} = z_{\text{QSO}},$$

that is, we insist the absorption center be within the range of the observed wavelengths (after restricting to the chosen domain $\lambda_{\text{rest}} \in [800\,\text{Å}, 1216\,\text{Å}]$). Given these, we simply take a uniform prior distribution on this range:

$$p(z_{\text{DLA}} \mid z_{\text{QSO}}, \mathcal{M}_{\text{DLA}}) = \mathcal{U}[z_{\min}, z_{\text{QSO}}].$$

The column density prior $p(N_{\text{HI}} \mid \mathcal{M}_{\text{DLA}})$ is slightly more complicated. Rather than selecting a parametric distribution for this prior, we make a nonparametric estimate of the density. Due to the large dynamic range of column densities, we chose a prior on its base-10 logarithm, $\log_{10} N_{\text{HI}}$. We use the reported values for the $N_{\text{DLA}} = 5\,854$ DLAs contained in the DR9 sample to make a kernel density estimate of the density $p(\log_{10} N_{\text{HI}} \mid \mathcal{M}_{\text{DLA}})$. Here we selected the univariate Gaussian probability density function for our kernels, with bandwidth selected via a plug-in estimator that is asymptotically optimal for normal densities.

4.3. Approximating the Model Evidence

Given our choice of parameter prior, the integral in (6) is unfortunately intractable, so we resort to numerical integration. In particular, we use quasi-Monte Carlo (QMC) integration (Caflisch, 1998), taking $N = 1\,000$ samples generated from a scrambled Halton sequence (Kocis & Whiten, 1997) to define our parameter samples. Note that the Halton sequence gives values approximately evenly distributed on the unit square $[0, 1]^2$, which (after a trivial transformation) agrees in density with our redshift prior, but not our column density prior. To correct for this, we used inverse transform sampling to endow the generated samples with the appropriate distribution via the approximate inverse cumulative distribution function implied by our kernel density estimate of $p(\log_{10} N_{\text{HI}} \mid \mathcal{M}_{\text{DLA}})$.

5. Model Prior

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

Only approximately 10% of the QSO sightlines in the DR9 and DR10 releases contain DLAs. A simple approach would be to use a fixed value of $\Pr(\mathcal{M}_{\text{DLA}}) \approx 1/10$. However, we are less likely to observe a DLA in low-redshift QSOs due to the wavelength coverage of the SDSS and BOSS spectrographs being limited to $\lambda_{\text{obs}} \geq 3\,800\,\text{Å}$ and $\lambda_{\text{obs}} \geq 3\,650\,\text{Å}$, respectively, on the blue end. Therefore, here we will use a slightly more sophisticated approach and derive a redshift-dependent model prior $\Pr(\mathcal{M}_{\text{DLA}} \mid z_{\text{QSO}})$.

Our prior is simple and data driven. Consider a QSO with redshift z_{QSO}. Let N be the number of QSOs in the training sample with redshift less than $z_{\text{QSO}} + z'$, where z' is a small constant; here we took $z' = 0.1$. Let M be the number of the sightlines of these containing DLAs. We define

$$\Pr(\mathcal{M}_{\text{DLA}} \mid z_{\text{QSO}}) = \tfrac{M}{N}.$$

The constant z' serves to ensure that QSOs with very small redshift have sufficient data for estimating the prior. The resulting prior ranges from roughly 2% for the lowest-redshift quasars to around 10% for high-redshift objects.

6. Experiment

We have now developed all the 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 Figure 2.

Consider a quasar with redshift z_{QSO}, and suppose we have made spectroscopic observations of the object $\mathcal{D} = (\boldsymbol{\lambda}, \mathbf{y})$, with known observation noise variance vector $\boldsymbol{\nu}$. First, we compute the prior probability of the DLA model $\mathcal{M}_{\text{DLA}}$, $\Pr(\mathcal{M}_{\text{DLA}} \mid z_{\text{QSO}})$ and thus the prior odds in favor of the DLA model; for our example, $\Pr(\mathcal{M}_{\text{DLA}} \mid z_{\text{QSO}}) = 10.3\%$, giving prior odds of 0.115 (9-to-1 against the DLA model).

Next, we compute the Bayes factor in favor of $\mathcal{M}_{\text{DLA}}$:

$$\frac{p(\mathbf{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}, \mathcal{M}_{\text{DLA}})}{p(\mathbf{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}, \mathcal{M}_{\neg\text{DLA}})}. \tag{7}$$

See (5) for how to compute the model evidence for the null model and Section 4.3 for our approximation to the DLA model evidence. For our example, the Bayes factor overwhelmingly supports the DLA model, with a value of $\exp(136) \approx 10^{59}$. The computation of the Bayes factor is illustrated in Figure 5.

Finally, the posterior odds in favor of the sightline containing an intervening DLA is the product of the prior odds and the Bayes factor (7). The log odds in favor of $\mathcal{M}_{\text{DLA}}$ for the example from Figure 2 are 134 nats, and the probability of the sightline containing a DLA is effectively unity. The DLA parameter sample with the highest likelihood was $(z_{\text{DLA}}, \log_{10} N_{\text{HI}}) = (3.286, 20.30)$, closely matching the values reported in the DLA catalog, $(z_{\text{DLA}}, \log_{10} N_{\text{HI}}) = (3.284, 20.29)$.

We note that despite the large number of parameter samples, our method is extremely scalable. The low-rank-plus-diagonal structure of our observation covariance ensures that fully processing a quasar spectrum takes under a second on a standard desktop machine.

To verify the validity of our proposed method, we computed the posterior odds in favor of $\mathcal{M}_{\text{DLA}}$ for each of the 101 167 quasar sightlines in the BOSS DR10 Lyman-α forest sample, using the GP emission model we built in Section 3 from the corresponding DR9 catalog.

To evaluate our results, we examined the ranking induced on the sightlines by the log posterior odds in favor of the DLA model. If our method is performing correctly, true DLAs should be at the top of this list, above the sightlines without DLAs. To visualize the quality of our ranking, we seek to create a receiver–operating characteristic (ROC) plot.

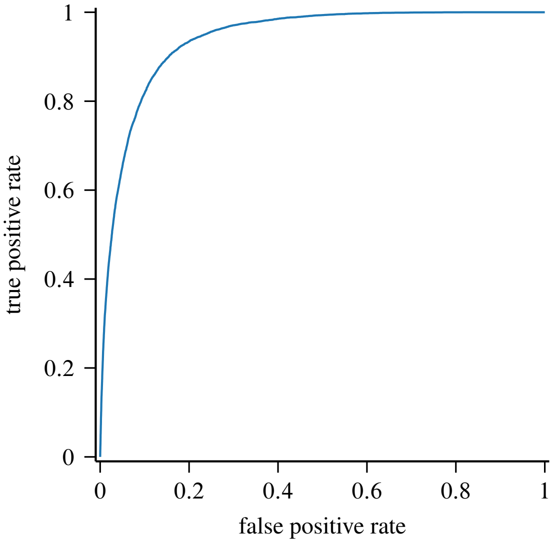
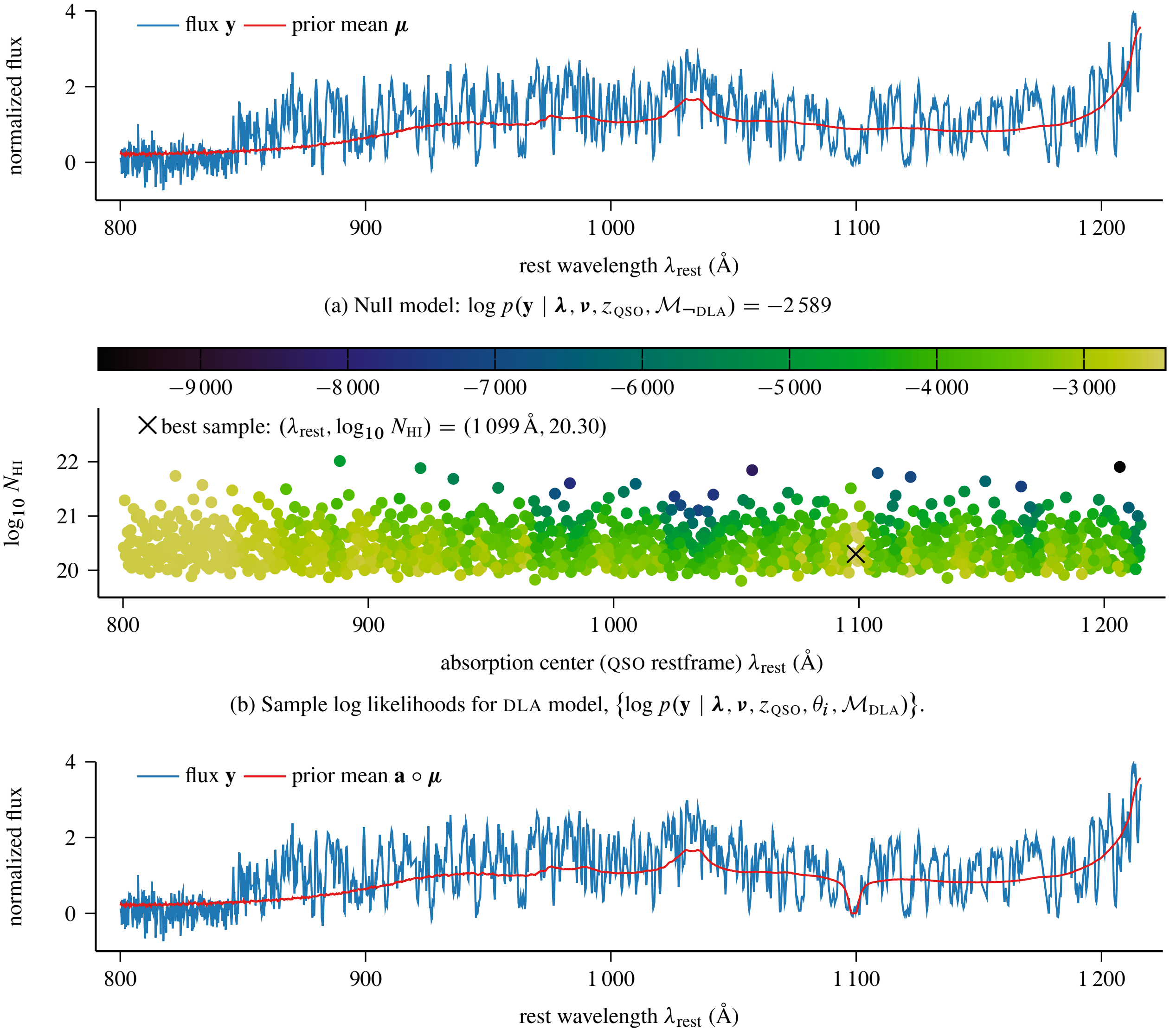


Figure 6. The ROC plot for the ranking of the 101 167 QSO sightlines contained in the BOSS DR10 Lyman-α forest sample induced by the log posterior odds of containing a DLA. The area under the curve is 94.07%.

Creating an ROC plot requires knowledge of the ground-truth labels for each of our objects, which of course we do not have. For this reason, we used the DR10 DLA concordance as a surrogate. The ROC plot for this surrogate is shown in Figure 6. The area under the ROC curve (AUC) statistic was 94.07%. The AUC is equivalent to the probability that a positive example, chosen uniformly at random, will be ranked higher than a negative example. Approximately 82% of the DLAs flagged in the catalog appear in the top-10% of the list. Clearly our approach is effective at identifying DLAs. We also note that if we restrict to only objects not appearing in the DR9 training sample, the AUC remains nearly constant at 94.03%, so there does not seem to be significant bias due to inclusion of some sightlines when building our null model.

An important caveat to these results is that the DLA concordance is unlikely to represent the absolute ground truth, and many “false positive” sightlines could in fact contain as-yet undiscovered DLAs. The conservative definition of the concordance catalog also suggests many true DLAs are not flagged. A visual scan of a sample of our “false positives” by domain experts suggested this is likely the case. Many of these have low signal-to-noise ratio and are likely to have been rejected by previous automated attempts. By carefully modeling both flux correlations and heteroskedastic measurement noise, our GP approach appears to be more robust than previous methods.



(a) Null model: $\log p(\mathbf{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}, \mathcal{M}_{\neg\text{DLA}}) = -2\,589$

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

(c) DLA model: $\log p(\mathbf{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}, \mathcal{M}_{\text{DLA}}) = -2\,453$.

Figure 5. An illustration of the proposed DLA-finding procedure for the quasar sightline in Figure 2. (a) shows the normalized flux with the prior GP mean for our learned null model $\mathcal{M}_{\neg\text{DLA}}$. (b) shows the log likelihoods for each of the parameter samples used to approximate the marginal likelihood of our DLA model $\mathcal{M}_{\text{DLA}}$. (c) shows the normalized flux with the prior GP mean associated with the best DLA sample, $(z_{\text{DLA}}, \log_{10} N_{\text{HI}}) = (3.286, 20.30)$, corresponding to $\lambda_{\text{obs}} = 1\,099$ Å.

7. Discussion

This work represents the first fully automated approach for detecting DLAs in large-scale surveys of quasar spectra, filling a critical need of the astronomical community. Our method is highly efficient, and will scale easily to upcoming massive surveys such as SDSS–IV.

We plan to compile a catalog of our results using the recently released, final SDSS–III data release, DR12. We also plan to make the parameters of our learned GP model, using the full DR10 release, available for future research efforts.

We regard this collaboration as a successful application of modern machine-learning methods to the myriad large-scale data processing issues faced by modern astronomy, and hope that techniques similar to those used here can be the basis for future cooperation.

Acknowledgments

Part of this research was supported under DOE grant number DESC0011114.

References

Caflisch, R. E. Monte Carlo and quasi-Monte Carlo methods. *Acta Numerica*, 7:1–49, 1998.

Carithers, W. C. DLA Concordance Catalog. Published internally to SDSS, 2012.

Draine, B. T. *Physics of the Interstellar and Intergalactic Medium*. Princeton University Press, 2011.

Eisenstein, D. J., Weinberg, D. H., Agol, E., et al. SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way, and Extra-Solar Planetary Systems. *The Astronomical Journal*, 142:72, 2011. doi: 10.1088/0004-6256/142/3/72.

Kocis, L. and Whiten, W. J. Computational Investigations of Low-Discrepancy Sequences. *ACM Transactions on Mathematical Software*, 23(2):266–294, 1997.

Lee, K.-G., Bailey, S., Bartsch, L. E., et al. The BOSS Lyα Forest Sample from SDSS Data Release 9. *The Astronomical Journal*, 145:69, 2013. doi: 10.1088/0004-6256/145/3/69.

Lee, K.-G. et al. The BOSS Lyα Forest Sample from SDSS Data Release 10. In preparation, 2014.

Noterdaeme, P., Petitjean, P., Carithers, W. C., et al. Column density distribution and cosmological mass density of neutral gas: Sloan Digital Sky Survey-III Data Release 9. *Astronomy and Astrophysics*, 547:L1, 2012. doi: 10.1051/0004-6361/201220259.

Rasmussen, C. E. and Williams, Christopher K. I. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Slosar, A., Font–Ribera, A., Pieri, M. M., et al. The Lyman-α forest in three dimensions: measurements of large scale flux correlations from BOSS 1st-year data. *Journal of Cosmology and Astroparticle Physics*, 9:001, 2011. doi: 10.1088/1475-7516/2011/09/001.

Smee, S. A., Gunn, J. E., Uomoto, A., et al. The Multi-object, Fiber-fed Spectrographs for the Sloan Digital Sky Survey and the Baryon Oscillation Spectroscopic Survey. *The Astronomical Journal*, 146:32, 2013. doi: 10.1088/0004-6256/146/2/32.

Wolfe, A. M., Gawiser, E., and Prochaska, J. X. Damped Lyα Systems. *Annual Review of Astronomy and Astrophysics*, 43:861–918, 2005. doi: 10.1146/annurev.astro.42.053102.133950.

Wright, E. L. Lyman Alpha Forest, 2004. URL `http://www.astro.ucla.edu/~wright/Lyman-alpha-forest.html`.