Gaussian Processes for Identifying Damped Lyman-α Systems in Spectroscopic Surveys

Roman Garnett, Shirley Ho, and Jeff Schneider

Robotics Institute
Department of Physics
Carnegie Mellon University
Pittsburgh, PA 15213
{rgarnett, schneider}@cs.cmu.edu
shirleyh@andrew.cmu.edu

Abstract

We present an application of Gaussian processes to the problem of discovering “damped Lyman-α systems” (DLAs), dense regions of hydrogen gas that are likely the birthplace of galaxies like the Milky Way. We apply this method to a large spectroscopic survey of quasars from the Sloan Digital Sky Survey III (SDSS-III).

1 Introduction

How did our Milky Way galaxy form? What conditions was it born into? One astronomical method for approaching these centuries-old fundamental questions involves investigating probable proto-galaxies called “damped Lyman-α systems” (DLAs), regions of space dense with neutral hydrogen gas. DLAs are characterized by telltale absorption features they leave in spectroscopic readings of faraway objects such as quasi-stellar objects (QSOs or “quasars”). Astronomers believe that these systems are most likely the birthplaces of galaxies like our own, and achieving a better understanding of these systems and their distribution is a current astronomical challenge of extreme importance.

Despite the importance of understanding these systems, until very recently only a few hundred identified examples of DLA systems were known, all found by unassisted visual inspection. The advent of the Sloan Digital Sky Survey III (SDSS-III) [3, 2], which aims to survey a total of 160,000 quasars with redshift $z > 2$, represented an unprecedented opportunity to study DLA systems. Despite the massive amount of data being collected by SDSS-III, the importance of finding and cataloging DLAs has compelled astronomers to continue visual inspection of collected spectra. This process is both tedious and its success rate is difficult to quantify. Two simple automated systems have been proposed for finding likely DLA-containing quasar spectra [1]; however, these methods have high disagreement with both each other and with the results of visual surveys. A so-called “concordance” list has been prepared [1], identifying the SDSS-III quasars which all three methods agree contain a DLA. This can serve as a “ground truth” set for any proposed DLA-location methods.

Here we will describe a technique based on Gaussian processes (GPs) to locate these systems. The proposed method allows for sophisticated probabilistic analysis of the collected spectra and DLA-related parameters not possible with current techniques. Before we describe our procedure, we will briefly discuss DLAs.

2 Damped Lyman-α Systems

In spectroscopy, we measure the relative intensity of light over a given range of wavelengths while observing a faraway object. Without any intervening matter, we would observe the light exactly as it

1
was emitted, translated towards longer wavelengths due to redshift caused by the Doppler effect and the expansion of our universe. If another object occludes our view, however, it can interact with the emitted photons and alter our measurements in ways rendered predictable by quantum mechanics.

A particularly noteworthy interaction can occur with photons at wavelengths near $\lambda = 1216$ Å, corresponding to the “Lyman-α” emission line of hydrogen. Quasar spectra in particular will show large intensity at this wavelength, allowing for their redshift to be easily determined. If a cloud of gas containing neutral hydrogen lies between us and the quasar, it will absorb a fraction of the emitted photon, causing a corresponding reduction in flux. Figure 1 shows an illustration. As the light travels from a quasar to Earth, it passes through many gas clouds, each leaving a corresponding absorption profile. Their combined effects give rise to a “spiky” region known as the “Lyman-α forest.” If the gas region is dense enough (corresponding to a column density $N_\ell$ of over approximately $10^{20}$ atoms cm$^{-2}$), the corresponding dip will be so large that almost no intensity will be observed in the center of the absorption feature, and its sides will show characteristic “damping wings.” When these features are present, the region is deemed a damped Lyman-α system.

The emitted flux at wavelength $\lambda$, $f_{\text{QSO}}(\lambda)$, and the observed flux, $f_{\text{OBS}}(\lambda)$, are related by $f_{\text{OBS}}(\lambda) = \exp(-\tau(\lambda))f_{\text{QSO}}(\lambda)$, where $\tau$ represents a multiplicative dampening function. Modulo the effect of this dampening function, the relative flux of quasars at similar redshifts is usually similar. For a particular absorbing system, a theoretical physical analysis of the system allows us to calculate $\tau$ given the redshift of the system, the redshift of the quasar, and the column density of the gas. For DLAs, $\tau$ can be well approximated by a scaled Cauchy distribution centered at $1216(1 + z_{\text{DLA}})$ Å whose scale parameter is a function of $N_\ell$.

3 Identifying DLAs with Gaussian processes

Here we describe our approach for locating DLAs in spectroscopic readings. We will take a Bayesian nonparametric approach and model the unknown emitted intensity of the quasar using a Gaussian process (GP) prior. We will construct two different GP models, one corresponding to the presence of a DLA between Earth and the quasar, and one to the absence of a DLA; after estimating the marginal likelihood of our measurements under each of these models, we perform Bayesian model selection.

3.1 Gaussian process prior on the emitted flux

We begin by placing a Gaussian process prior distribution on the unknown spectrum emitted by the quasar, $f_{\text{QSO}}$:

$$p(f_{\text{QSO}}(\lambda) \mid \theta) = \mathcal{GP}(f_{\text{QSO}}(\lambda); \mu_{\text{QSO}}(\lambda; \theta), K_{\text{QSO}}(\lambda, \lambda'; \theta)),$$
Figure 2: The relative flux of a quasar ($z_{\text{QSO}} = 3.11$) over the range $[z_{\text{QSO}} - 1, z_{\text{QSO}}]$. The blue line shows the prior mean function of the maximum likelihood fit of the DLA GP model, corresponding to $(z_{\text{DLA}}, \log_{10} N_\ell) = (2.68, 20.68)$; the concordance lists $(z_{\text{DLA}}, \log_{10} N_\ell) = (2.68, 20.44)$. For this example, $\log p_{\text{DLA}} - \log p_{\text{DLA}} = 8.35$ nats.

where $\theta$ represents the vector of all hyperparameters. In this application, we selected the Matérn covariance function with $\nu = 3/2$ for $K_{\text{QSO}}$. For the prior mean, we selected

$$\mu_{\text{QSO}}(\lambda) \propto \hat{f}_{\text{QSO}}(\lambda; z_{\text{QSO}}),$$

where $\hat{f}$ is an empirically approximated mean normalized flux for quasars at a given redshift. Given a survey containing tens of thousands of quasar spectra such as SDSS-III, we may approximate $\hat{f}$ using fairly simple binning methods.

### 3.2 Implied Gaussian process prior on the observed flux

This prior on $f_{\text{QSO}}$ implies the following prior on $f_{\text{OBS}}$:

$$p(f_{\text{OBS}}(\lambda) \mid \theta) = GP (f_{\text{OBS}}(\lambda); a(\lambda) \mu_{\text{QSO}}(\lambda; \theta), a(\lambda) K_{\text{QSO}}(\lambda, \lambda'; \theta) a(\lambda'),$$

where $a(\lambda) = \exp(-\tau(\lambda))$. Unfortunately, there is no way for us to know the damping function exactly. In the absence of a DLA, we instead settle on setting $\tau = 0$ everywhere; in the presence of a DLA at a given redshift and column density, we set $\tau$ to the appropriate theoretical profile.

Finally, we assume a simple observation model, where our measurements are corrupted by zero-mean iid Gaussian noise.

### 3.3 Calculating marginal likelihoods

With our models specified, we now wish to calculate the probability of our observations under both of them, which we will denote with $p_{\sim \text{DLA}}$ and $p_{\text{DLA}}$.

Let our observations be denoted by $D = (\lambda, f)$. Calculating the required model likelihoods requires evaluating the nonanalytic integral

$$p(f \mid D) = \int p(f \mid \lambda, \theta) p(\theta) d\theta,$$

where we have marginalized the unknown hyperparameters.

In the non-DLA case, we have four hyperparameters: the scaling of $\hat{f}_{\text{QSO}}$ for the prior mean, the length and output scales for $K_{\text{QSO}}$, and the noise variance. In this case, our measurements are typically quite dense and the likelihood surface $p(\theta \mid D)$ is extremely peaked. In this case, we may approximate $p_{\sim \text{DLA}} = p(f \mid \lambda)$ with $p(f \mid \lambda, \theta_{\sim \text{DLA}-\text{ML}})$, where $\theta_{\sim \text{DLA}-\text{ML}}$ are the maximum likelihood hyperparameters in the non-DLA case. This was our approach.

In the DLA case, things are somewhat more complicated due to the parameters $z_{\text{DLA}}$ and $N_\ell$. The likelihood surface over these two parameters is typically not nearly as peaked as for the non-DLA
parameters above (which we continue to approximate with the same point estimate \( \theta \_{\text{DLA-ML}} \)). We therefore wish to estimate (1) with an appropriate prior distribution over the DLA hyperparameters. This inference problem has been studied in a Bayesian context by several authors under the name “Bayesian quadrature” or “Bayesian Monte Carlo;” see [4] for a discussion. This allows us to, if desired, also approximate the full posterior distributions of \( z_{\text{DLA}} \) and \( N_{\ell} \).

4 SDSS-III Experiment

We now briefly explain an experiment carried out on the SDSS-III data meant to demonstrate the technique described here. We collected spectra corresponding to 1 000 quasars from the DLA concordance list (serving as ground-truth “positive” examples) and 1 000 quasars not on the list (serving as ground-truth “negative” examples).

For each processed spectrum, we extracted the measurements corresponding to \( \lambda \in [z_{QSO} - 1, z_{QSO}] \); this interval accounted for 98\% of the DLAs in the concordance list. We used these data to calculate \( p_{DLA} \) using maximum likelihood as described above, and approximated \( p_{\neg DLA} \) using a simple Bayesian Monte Carlo estimate with independent Gaussian priors on \( \log(z_{QSO} - z_{DLA}) \) and \( \log_{10} N_{\ell} \), whose parameters were estimated from the concordance list [4].

Figure 2 shows an example, including the prior mean corresponding to the maximum likelihood parameters for the DLA model. The DLA model was favored considerably (\( \log p_{DLA} - \log p_{\neg DLA} = 8.35 \) nats), and the maximum likelihood parameters agree well with the previously reported values. Figure 3 (left) shows an approximation of the full posterior distribution over \( \log_{10} N_{\ell} \) at the most likely \( z_{DLA} \). Finally, Figure 3 (right) shows a receiver–operating characteristic (ROC) plot for the processed quasars ordered by \( \log p_{DLA} - \log p_{\neg DLA} \). The incredibly steep slope at low false positive rates suggest that this method could help considerably at prioritizing visual inspection. Notice that the curve shown is in fact probably lower than the true performance—some of the “false positives” might nonetheless contain DLAs.

References