The Astrophysical Journal, 864:143 (17pp), 2018 September 10 https://doi.org/10.3847/1538-4357/aad7f8

Predicting Quasar Continua near Lyα with Principal Component Analysis

Frederick B. Davies[1], Joseph F. Hennawi[1,2], Eduardo Bañados[3,9], Robert A. Simcoe[4], Roberto Decarli[2,5], Xiaohui Fan[6], Emanuele P. Farina[1,2], Chiara Mazzucchelli[2], Hans-Walter Rix[2], Bram P. Venemans[2], Fabian Walter[2], Feige Wang[1,7,8], and Jinyi Yang[6,7,8]

[1] Department of Physics, University of California, Santa Barbara, CA 93106-9530, USA; davies@physics.ucsb.edu
[2] Max Planck Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany
[3] The Observatories of the Carnegie Institution for Science, 813 Santa Barbara Street, Pasadena, CA 91101, USA
[4] MIT-Kavli Center for Astrophysics and Space Research, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
[5] INAF–Osservatorio Astronomico di Bologna, via Gobetti 93/3, I-40129 Bologna, Italy
[6] Steward Observatory, The University of Arizona, 933 North Cherry Avenue, Tucson, AZ 85721-0065, USA
[7] Department of Astronomy, School of Physics, Peking University, Beijing 100871, People's Republic of China
[8] Kavli Institute for Astronomy and Astrophysics, Peking University, Beijing 100871, People's Republic of China
Received 2018 January 22; revised 2018 July 10; accepted 2018 July 30; published 2018 September 11

Abstract

Measuring the proximity effect and the damping wing of intergalactic neutral hydrogen in quasar spectra during the epoch of reionization requires an estimate of the intrinsic continuum at rest-frame wavelengths of $\lambda_{\rm rest} \sim 1200$–$1260$ Å. In contrast to previous works which used composite spectra with matched spectral properties or explored correlations between parameters of broad emission lines, we opted for a nonparametric predictive approach based on principal component analysis (PCA) to predict the intrinsic spectrum from the spectral properties at redder (i.e., unabsorbed) wavelengths. We decomposed a sample of 12764 spectra of $z \sim 2$–2.5 quasars from the Sloan Digital Sky Survey (SDSS)/Baryon Oscillation Spectroscopic Survey (BOSS) into 10 red-side (1280 Å $< \lambda_{\rm rest} < 2900$ Å) and 6 blue-side (1180 Å $< \lambda_{\rm rest} < 1280$ Å) PCA basis spectra, and constructed a projection matrix to predict the blue-side coefficients from a fit to the red-side spectrum. We found that our method predicts the blue-side continuum with $\sim$6%–12% precision and $\lesssim$1% bias by testing on the full training set sample. We then computed predictions for the blue-side continua of the two quasars currently known at $z > 7$: ULAS J1120+0641 ($z = 7.09$) and ULAS J1342+0928 ($z = 7.54$). Both of these quasars are known to exhibit extreme emission line properties, so we individually calibrated the uncertainty in the continuum predictions from similar quasars in the training set, finding comparable precision but moderately higher bias than the predictions for the training set as a whole, although they may face additional systematic uncertainties due to calibration artifacts present in near-infrared echelle spectra. We find that both $z > 7$ quasars, and in particular ULAS J1342+0928, show signs of damping wing-like absorption at wavelengths redward of Lyα.

Key words: cosmology: observations – methods: numerical – quasars: emission lines – quasars: general

1. Introduction

The damping wing of neutral hydrogen Lyα absorption in the intergalactic medium (IGM) is predicted to be a key signature of the epoch of reionization at $z > 6$ (Miralda-Escudé 1998). This damped absorption signature should be very broad, affecting rest-frame wavelengths redward of Lyα ($\lambda_{\rm rest} = 1215.67$ Å) out to $\lambda_{\rm rest} \sim 1260$ Å if the IGM is mostly neutral. Measurement of this signal, however, requires knowledge of the intrinsic (i.e., unabsorbed) profile of the quasar spectrum, which in the wavelength range relevant to the IGM damping wing consists of a combination of multiple Lyα and N V broad emission line components in addition to a smooth underlying continuum. Our ability to measure the IGM damping wing is thus mostly limited by our ability to predict the shape of this part of the quasar spectrum.

Lacking an accurate physical model to predict the combined emission from the quasar accretion disk and the broad-line region, we must instead resort to an empirical approach, perhaps aided by machine learning. Fortunately, thousands of quasars have been observed by the Sloan Digital Sky Survey (SDSS), and these spectra hold a wealth of information relating the correlated strengths and properties of their broad emission lines. The challenge lies in how exactly to extract these correlations quantitatively to predict one part of the spectrum from measurements of a different part. In this case, we would like to predict the spectral region potentially affected by IGM absorption ($\lambda_{\rm rest} < 1280$ Å, henceforth the blue side of the spectrum) from the remaining redward spectral coverage ($\lambda_{\rm rest} > 1280$ Å, henceforth the red side of the spectrum).

Correlations between various spectral features in the spectra of quasars have been studied for decades (e.g., Boroson & Green 1992), and strong correlations are known to exist between various broad emission lines from the rest-frame ultraviolet to the optical (e.g., Shang et al. 2007). In principle, then, it should be possible to use the information contained in the red-side portion of the quasar spectrum to predict the quasar continuum on the blue side. Various techniques exist in the literature for predicting the quasar continuum close to Lyα, including the direct approach of Gaussian fitting the red side of the Lyα line (e.g., Kramer & Haiman 2009) and stacking of quasar spectra with similar (non-Lyα) emission line properties (e.g., Mortlock et al. 2011; Simcoe et al. 2012; Bañados et al. 2018). The most sophisticated predictive model to date was presented by Greig et al. (2017b), who determined covariant relationships between parameters of Gaussian fits to broad emission lines of C IV, C III], and S IV+O IV] and those of Gaussian fits to the Lyα line.

[9] Carnegie-Princeton Fellow.

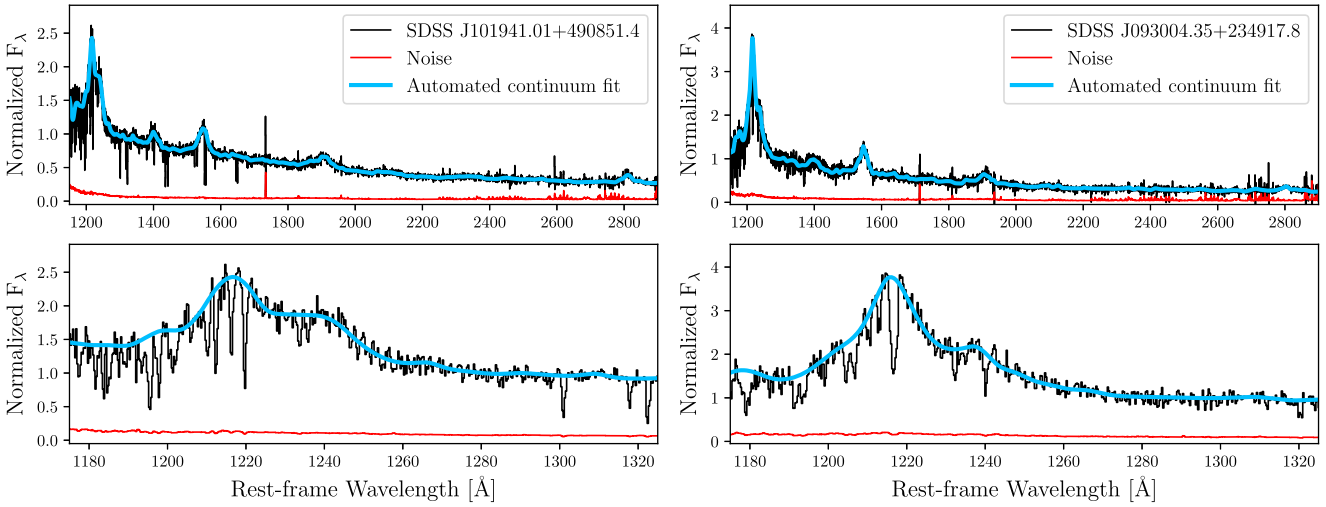

Figure 1. SDSS DR12 spectra of two quasars (black), their noise vectors (red), and their associated auto-fit continua (blue). The wavelength axes have been normalized to each quasar's rest frame, and the flux densities have been normalized to unity at $\lambda_{\rm rest} = 1290$ Å. In addition to the Lyα+N V+Si II complex at the blue end of the spectra, prominent broad emission lines of Si IV, C IV, C III], and Mg II are visible.

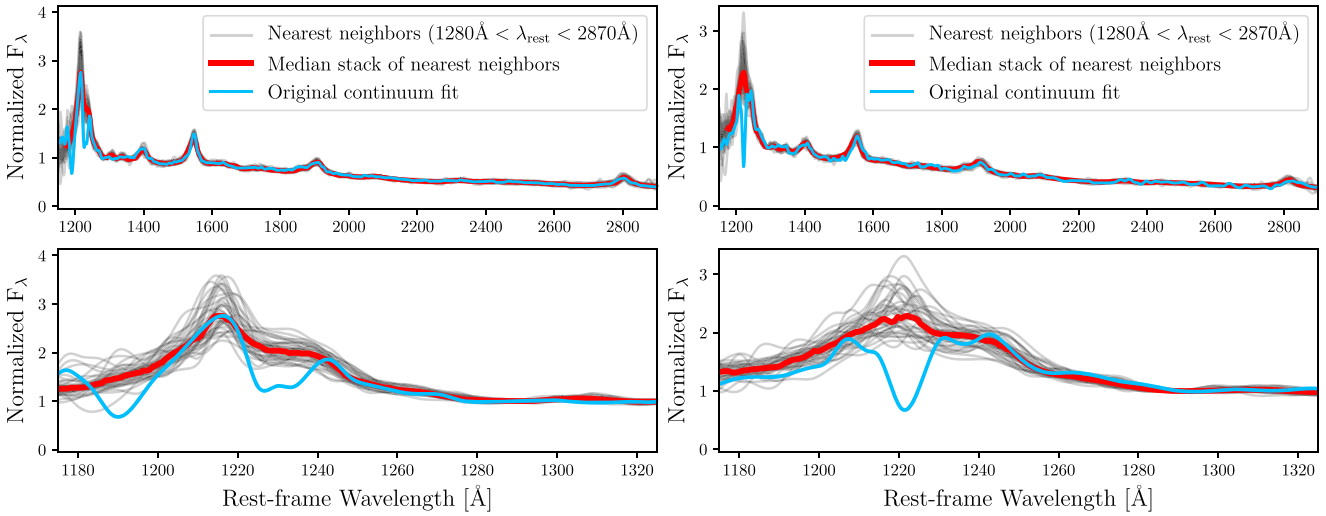

Figure 2. Auto-fit continua (blue) of SDSS J000113.15+322331.8 (left) and SDSS J235144.37+085649.4 (right), their 40 nearest-neighbor spectra for wavelengths 1280 Å $< \lambda_{\rm rest} <$ 2870 Å (gray), and the median stacks of those nearest neighbors (red). The strong associated absorption in the original spectra, shown more clearly in the bottom panels, is no longer present in the nearest-neighbor stacks.

Complicating matters is the fact that the spectral properties of quasars at $z \gtrsim 6.5$—most notably the properties of their C IV broad emission lines—appear to preferentially occupy a sparsely populated tail in the distribution of lower redshift quasars (Mazzucchelli et al. 2017). The two quasars known at $z > 7$, and in particular, the newly discovered highest redshift quasar ULAS J1342+0928 (Bañados et al. 2018), show very large C IV blueshifts relative to the general quasar population. The blueshift of the C IV emission line is correlated with properties of the Lyα emission line (e.g., Richards et al. 2011), and thus must be properly accounted for when modeling the Lyα region of high-redshift quasars (Bosman & Becker 2015). Any method trained on typical low-redshift quasars must then be able to perform well on objects which lie in the tails of the distribution of spectral properties.[10]

[10] Another option would be to simply restrict the training set to quasars with similar red-side properties, however there may be too few of such analogs (e.g., the 46 C IV-based analogs of ULAS J1342+0928 found by Bañados et al. 2018) to build a flexible model.

A common nonparametric method for determining correlations between different regions of quasar spectra is principal component analysis (PCA), wherein a set of input spectra is decomposed into eigenspectra that correspond to modes of common variations between different spectra. PCA was first applied to the spectral properties of quasars by Boroson & Green (1992) through analysis of not the spectral pixels themselves, but of the properties of various emission lines in the rest-frame optical. Francis et al. (1992) were the first to apply PCA to the spectral pixels themselves, using a relatively small sample of rest-frame UV quasar spectra to investigate PCA as a tool for quasar classification (see also Yip et al. 2004; Suzuki 2006). The idea of using PCA to predict the intrinsic continuum of absorbed regions of the spectrum was introduced by Suzuki et al. (2005), who constructed a predictive PCA model from low-redshift ($z \sim 0.14$–1.04) quasars in the context of predicting the unabsorbed continuum in the Lyα forest of higher redshift quasars where the continuum level cannot be measured directly. This technique was revisited by Pâris et al. (2011) with a somewhat larger

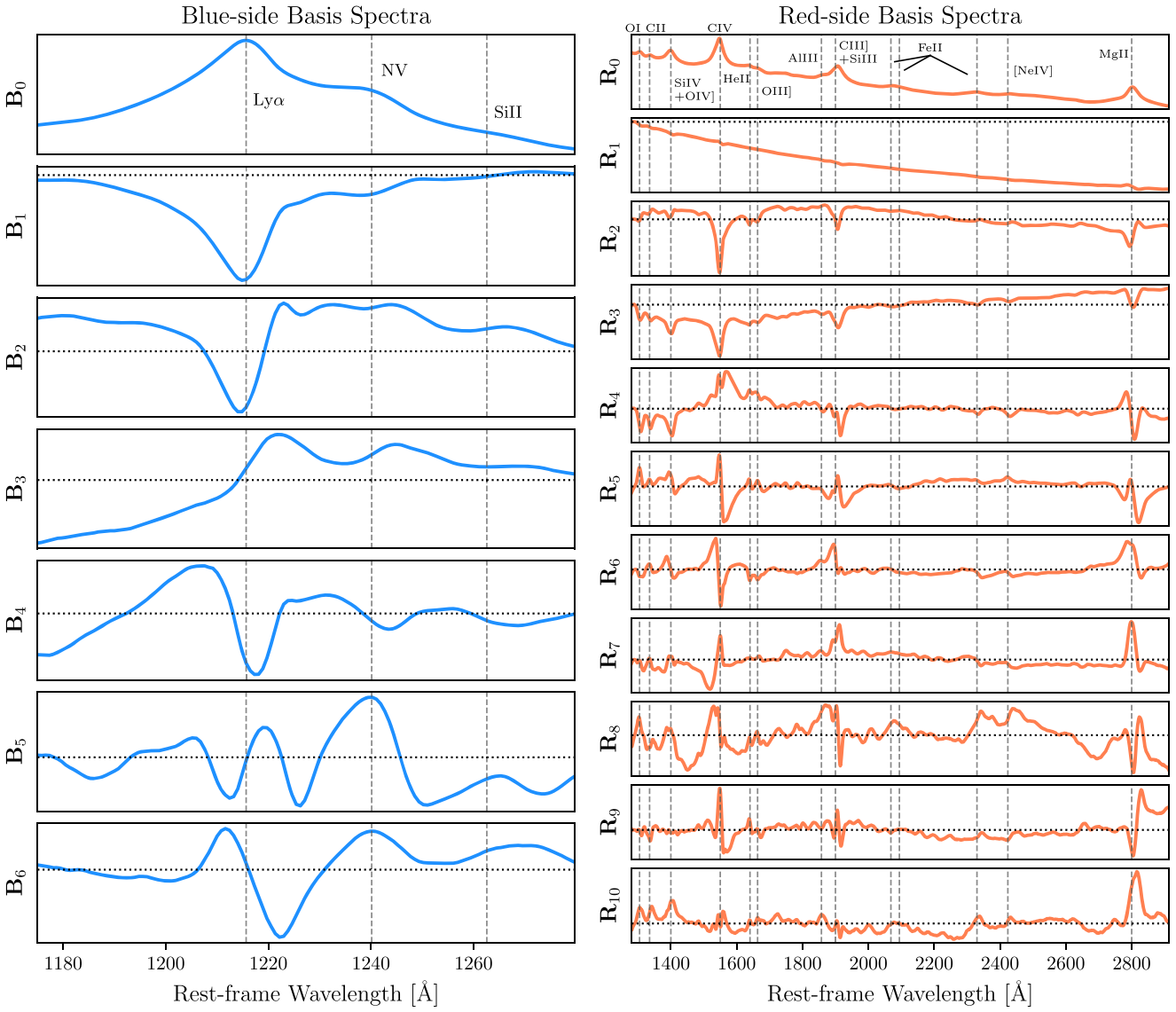

Figure 3. Blue-side and red-side basis spectra derived from the log-space PCA decomposition of 12,764 SDSS DR12 quasar spectra. The "0th" component in the top panels represents the mean of the log of the spectra, while the lower panels show the basis spectra ordered from highest to lowest variance explained from top to bottom. Vertical dashed lines highlight the central wavelengths of transitions of various species (or average wavelengths, in the case of blends) corresponding to broad emission lines in the spectrum, with line identifications taken from the SDSS/BOSS composite spectrum of Harris et al. (2016). The horizontal dotted line in each panel represents the zero level.

sample of high signal-to-noise (S/N) spectra of $z \sim 3$ quasars to estimate the evolution of the Lyα forest mean flux. Example applications of these methods include Lee et al. (2012) who applied the PCA models of Suzuki et al. (2005) and Pâris et al. (2011) to estimate the continuum level in the Lyα forest, and Eilers et al. (2017) who used the same PCA models to measure the sizes of proximity zones in $z \sim 6$ quasar spectra. These PCA models, however, were built from small ($N_{\rm qso} < 100$) samples of quasar spectra, and thus they are not well suited to reconstruct the spectra of outliers, e.g., the quasars known at $z \gtrsim 6.5$ with large emission line blueshifts.

In this work, we develop a PCA-based model for predicting the blue-side quasar continuum from the red-side spectrum in a similar vein to Suzuki et al. (2005) and Pâris et al. (2011), but from a much larger sample of spectra encompassing a wide range of spectral properties. In Section 2 we construct the training set of quasar spectra that serves as the foundation of our predictive model. In Section 3 we compute a set of basis spectra via a PCA decomposition of the spectra (in log space) and calibrate a relationship (i.e., a projection matrix) between red-side and blue-side PCA coefficients. In Section 4 we test the predictive power of the projections from the red side to the blue side on the training set spectra, and quantify the resulting (weak) bias and covariant uncertainty of the predicted blue-side continua. In Section 5 we apply our predictive model to the two quasars known at $z > 7$, which appear to show signs of damped Lyα absorption from the IGM. Finally, in Section 6 we conclude with a summary and describe future applications of this continuum model.

2. Definition of the Training Set

We draw our training set of quasar spectra from the SDSS-III/Baryon Oscillation Spectroscopic Survey (BOSS; Eisenstein et al. 2011; Dawson et al. 2013), which obtained

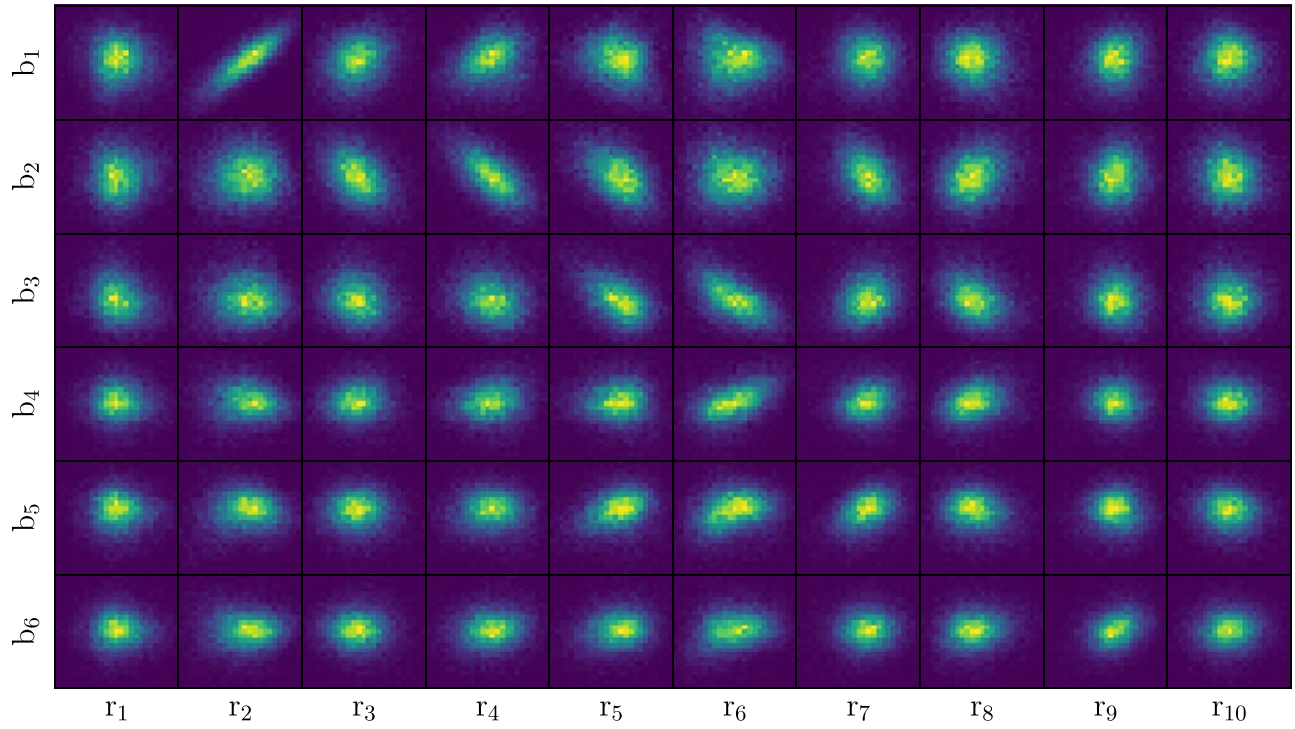

Figure 4. Joint distributions of best-fit red-side (horizontal axis) and blue-side (vertical axis) PCA coefficients from the training set. Strong and weak (anti-) correlations are visible between several of the coefficients; it is the information in these correlations that allow us to predict the blue-side spectrum from the red side.

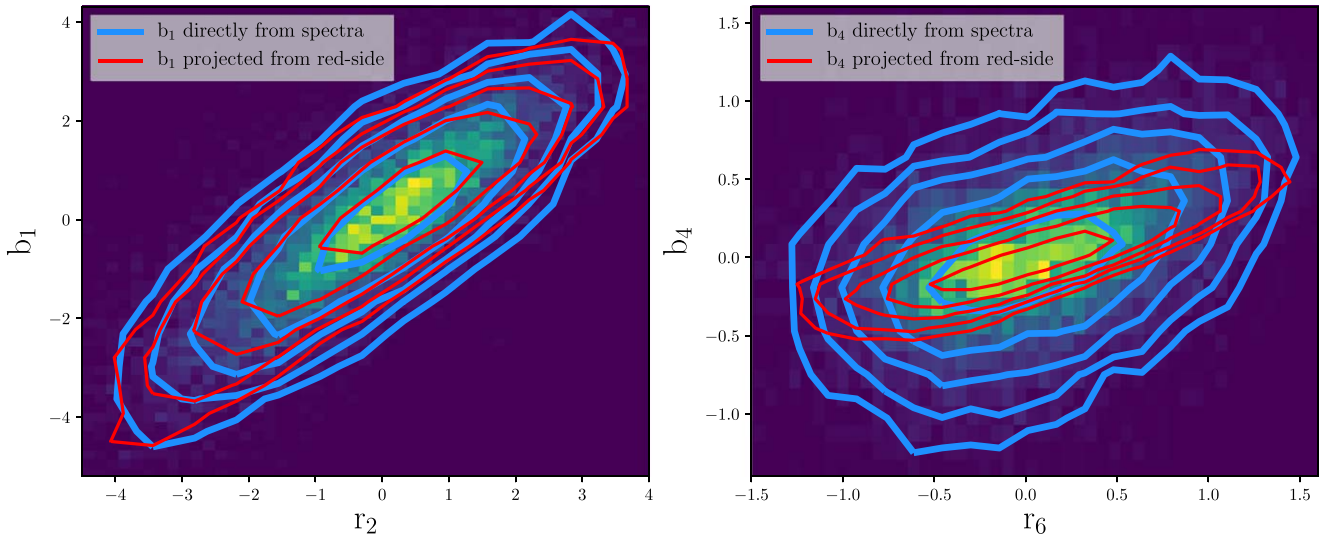

Figure 5. Left: distribution of r_2 and b_1 coefficients determined for the nearest-neighbor-stacked training set spectra (2D histogram and blue contours), as discussed in Section 3. The red contours show the distribution of r_2 and projected b_1 after applying Equation (5) to the full set of r_i for each spectrum. Right: same as the left panel, but for r_6 and b_4. The excess scatter in the blue-side coefficients determined from the spectra relative to the projections may reflect stochasticity in the relationship between the red-side and blue-side spectral features, or a nonlinear relationship that is not accounted for in the projection matrix (Equation (5)).

$R \sim 1200$–2500 spectra covering $\lambda_{\rm obs} \sim 3560$–$10400$ Å (Smee et al. 2013) of 294,512 quasars. We make use of the SDSS DR12 (Alam et al. 2015) quasar catalog (Pâris et al. 2017) with a query to the quasar spectrum database IGMSpec (Prochaska 2017) for quasars with BOSS pipeline redshifts of $2.09 < z_{\rm pipe} < 2.51$,[11] `BAL_FLAG_VI` $= 0$ to reject broad absorption line (BAL) quasars, and `ZWARNING` $= 0$ to avoid objects with highly uncertain redshifts. This redshift range was chosen for similar reasons as Greig et al. (2017b): the spectra cover the Lyα and Mg II broad emission lines.

We then compute the median S/N at $\lambda_{\rm rest} = \lambda_{\rm obs}/(1 + z_{\rm pipe}) = 1290 \pm 2.5$ Å for each spectrum, and apply an S/N cut of >7.0. This selection resulted in 13,328 quasars with complete wavelength coverage from $\lambda_{\rm rest} \sim 1170$–$2900$ Å, covering a similar range as observed for $z > 7$ quasars, and a median S/N of 10.1 at $\lambda_{\rm rest} = 1290 \pm 2.5$ Å. Further references to the S/N of the spectra will refer to the S/N in this wavelength range if not otherwise specified. Each spectrum was then normalized such that its median flux at $\lambda_{\rm rest} = 1290 \pm 2.5$ Å is unity.

[11] Motivated by the Appendix of Greig et al. (2017b), we use $z_{\rm pipe}$ rather than $z_{\rm PCA}$ because the C IV emission line shifts do not show redshift dependencies due to sky lines.

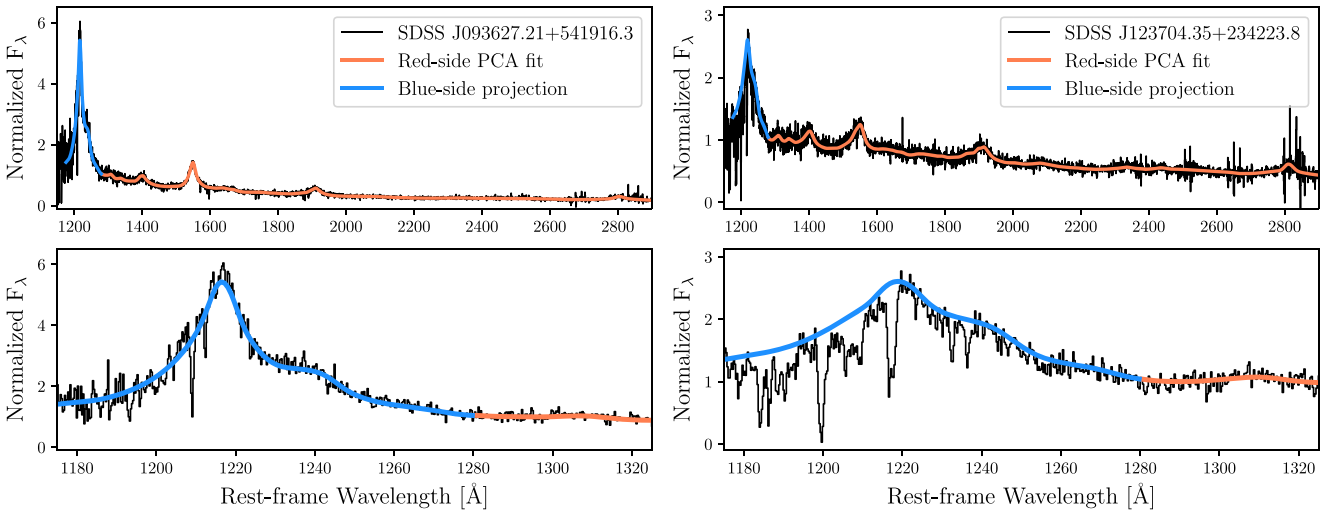

Figure 6. Example red-side PCA fits (orange) and blue-side predictions (blue) of two SDSS/BOSS quasar spectra (black) with S/N $\sim$ 10 at $\lambda_{\rm rest} = 1290$ Å. The top panels show the entire spectrum, while the bottom panels focus on the Lyα region.

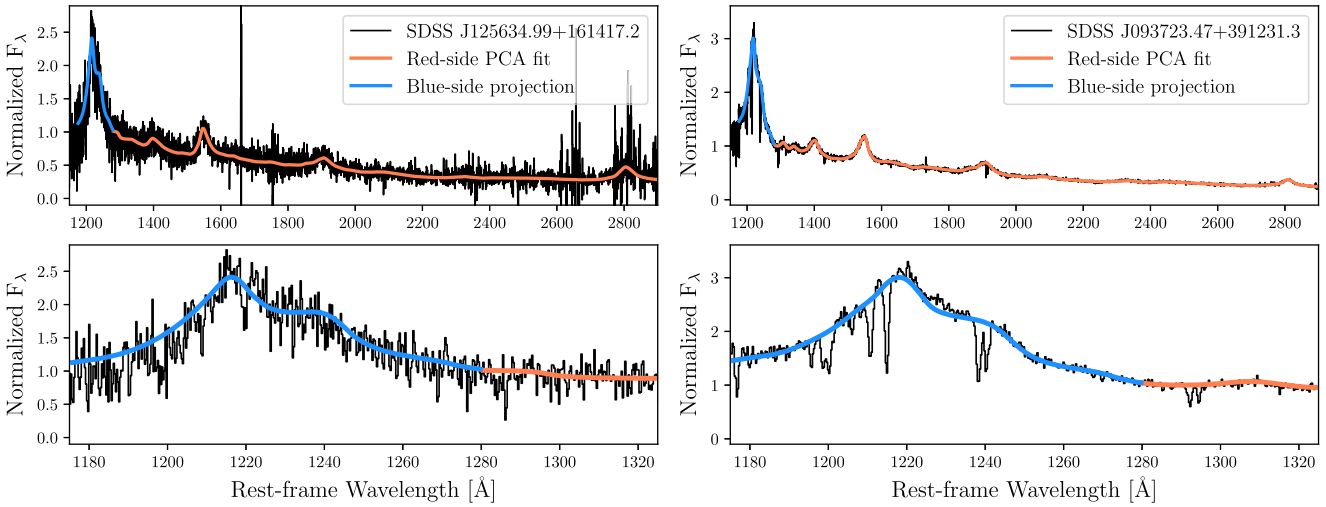

Figure 7. Same as Figure 6, but for example quasar spectra with S/N $\sim$ 7 (left) and S/N $\sim$ 25 (right) at $\lambda_{\rm rest} = 1290$ Å.

We then applied an automated continuum-fitting method developed by Young et al. (1979) and Carswell et al. (1982), as implemented by Dall'Aglio et al. (2008), which is designed to recover a smooth continuum in the presence of absorption lines both inside and outside of the Lyα forest. In brief, the method consists of dividing the spectra into 16 pixel segments ($\sim$1100 km s^{-1}), fitting continuous cubic spline functions to each segment, and iteratively rejecting pixels in each segment which lie more than two standard deviations below the fit (i.e., pixels inside of absorption lines).

We show two examples of quasar spectra (S/N $\sim$ 12) and their continuum fits in Figure 1. At this stage, we identify spectra whose continuum fits (normalized to be unity at $\lambda_{\rm rest} = 1290$ Å) fall below 0.5 at $\lambda_{\rm rest} < 1280$ Å and remove them from the analysis—these 357 quasars exhibit the strongest associated absorption (e.g., damped Lyα absorbers, strong N V absorption) or are remaining BAL quasars which were not flagged by the visual inspection of Pâris et al. (2017). We additionally remove 207 quasars whose continua drop below 0.1 at $\lambda_{\rm rest} > 1280$ Å, because such a weak continuum relative to $\lambda_{\rm rest} = 1290$ Å is indicative of very low S/N in the red-side spectrum, which typically implies significant systematics from OH line sky-subtraction residuals. Applying all of these criteria, our final training set consists of the remaining 12,764 quasar spectra.

Our sample of auto-fit quasar continua still contains a small fraction of "junk," typically quasars with strong associated absorption that were not caught by the blue-side criterion above or other artifacts that are not straightforward to eliminate in an automated fashion. To further clean up the training set in an objective manner, we replace each spectrum with a median stack of the original spectrum and its 40 nearest neighbors in the set of auto-fit continua, where the neighbors have been defined via a Euclidean distance in (normalized) flux units. That is,

$$D_{ij} = \sqrt{\sum_{\lambda} (C_{\lambda,i} - C_{\lambda,j})^2}, \quad (1)$$

where i and j denote two different quasar spectra and C_λ is the normalized auto-fit continuum. To avoid combining spectra with similar associated absorbers present in the Lyα+N V region, we find the nearest neighbors only using pixels with $\lambda_{\rm rest} > 1280$ Å. In Figure 2 we show the resulting nearest-

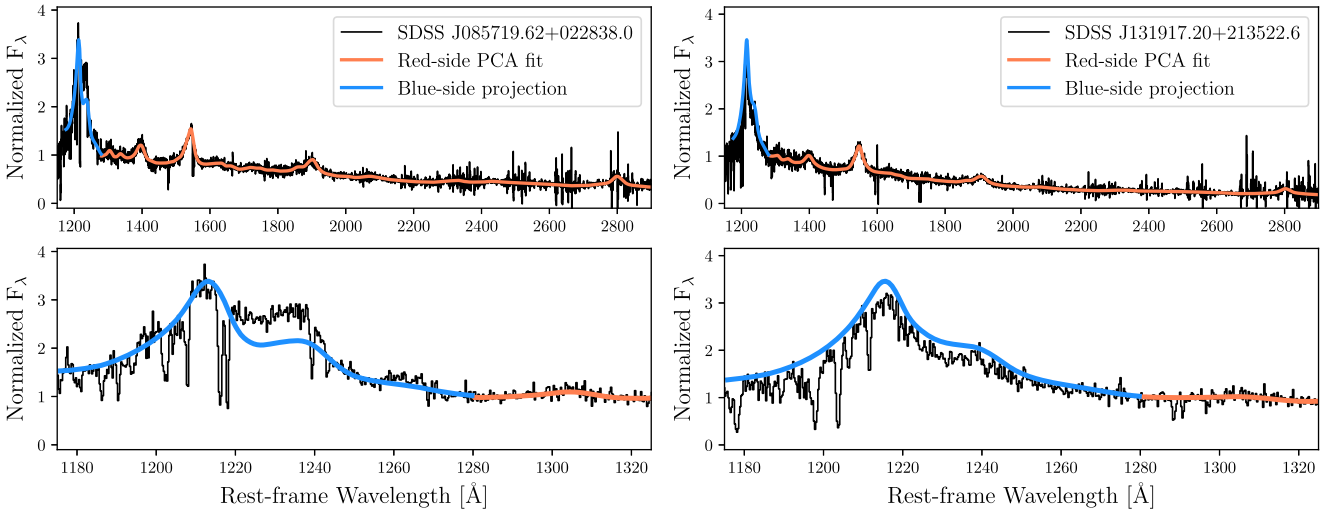

Figure 8. Same as Figure 6, but for quasar spectra with relatively poor blue-side predictions. In the left panel we show a prediction that undershoots the true continuum, and in the right panel we show a prediction that overshoots the true continuum.

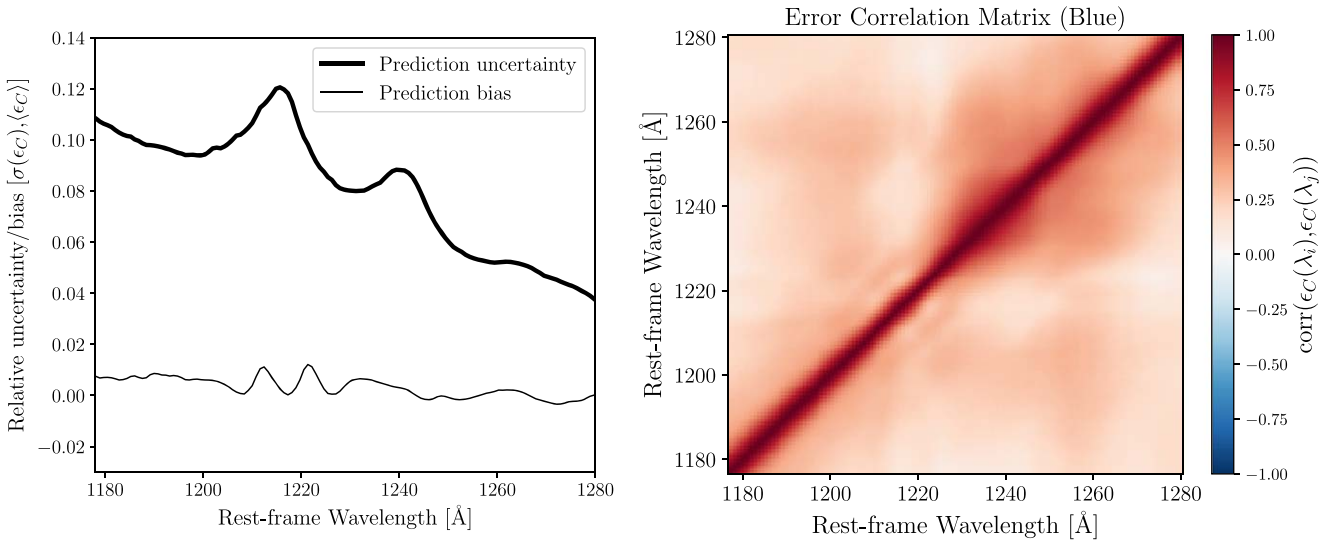

Figure 9. Left: relative uncertainty (1σ, thick curve) and mean bias (thin curve) of the blue-side PCA predictions applied to the full training set sample. Locations of broad emission lines of Lyα and N V appear as spectral regions with increased uncertainty. The most important wavelength range for constraints from the proximity zone and IGM damping wing is $\lambda_{\rm rest} \sim 1210$–$1250$ Å. Right: correlation matrix of PCA blue-side prediction errors. The blurring along the diagonal is due to the scale of the spline fit continua, while the larger scale correlations are due to errors in matching broad emission line features.

neighbor stacks (red) for two S/N $\sim$ 11 quasars with strong associated absorbers, seen as the large dips in the original continuum fit (cyan) close to Lyα, whose spectra would otherwise contribute unwanted features (i.e., not intrinsic quasar emission) to the analysis.

3. Log-space PCA Decomposition and Blue-side Projection

The PCA decomposes a set of training spectra into a set of orthogonal basis spectra, such that a spectrum can be expressed as

$$\boldsymbol{F} \approx \langle \boldsymbol{F} \rangle + \sum_{i=1}^{N_{\rm PCA}} a_i \boldsymbol{A}_i \,, \tag{2}$$

where a_i are the coefficients associated with each basis spectrum $\boldsymbol{A}_i$, and $N_{\rm PCA}$ is the number of basis spectra used for the reconstruction. PCA basis spectra represent the dominant modes of variance between spectra in the training set, in order from most to least amount of variance explained. The dominant mode of variation between quasar spectra, however, is the varying slope of the (roughly) power-law continuum. Power-law variations are not naturally described by additive components (e.g., Lee et al. 2012), but they are perfectly described by a single multiplicative component, i.e., $F_\lambda = \langle F_\lambda \rangle \times \lambda^{\Delta\alpha}$ where $\langle F_\lambda \rangle$ is the average quasar spectrum and $\Delta\alpha$ is a change in spectral index. Motivated by this, we perform the PCA decomposition in *log space*,

$$\log \boldsymbol{F} \approx \langle \log \boldsymbol{F} \rangle + \sum_{i=1}^{N_{\rm PCA}} a_i \boldsymbol{A}_i \,, \tag{3}$$

or in other words,

$$\boldsymbol{F} \approx e^{\langle \log \boldsymbol{F} \rangle} \times \prod_{i=1}^{N_{\rm PCA}} e^{a_i \boldsymbol{A}_i}. \tag{4}$$

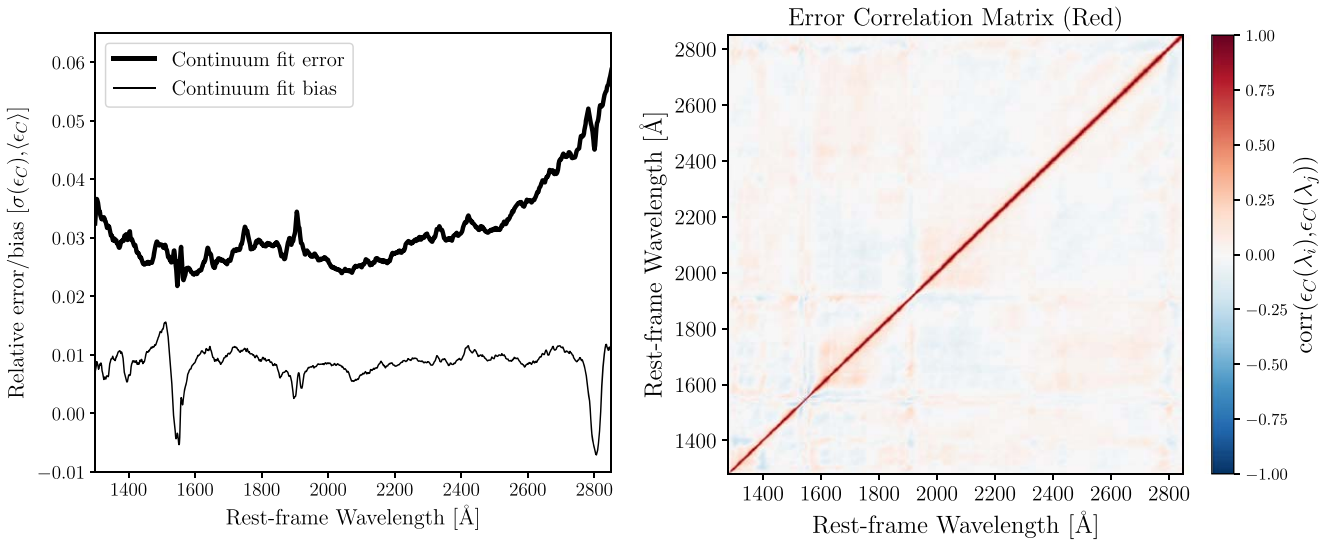

Figure 10. Left: relative error (1σ, thick curve) and mean bias (thin curve) of the red-side PCA continuum fits of the full training set sample. Right: correlation matrix of red-side PCA continuum fit errors.

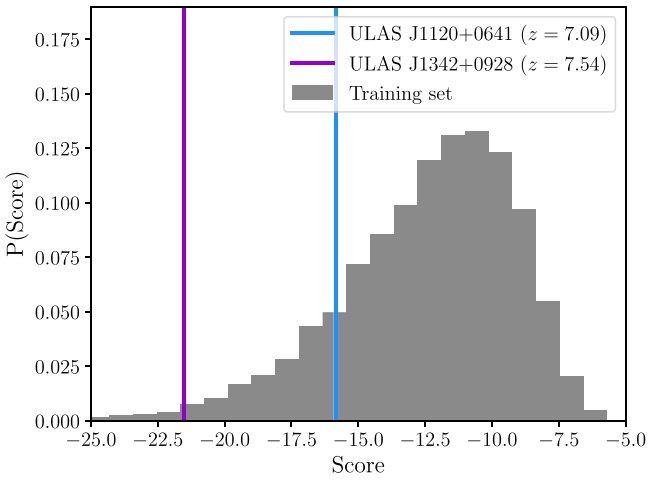

Figure 11. Distribution of score—log of GMM probability (see Section 5)—for the best-fit red-side coefficients of the training set quasars (histogram). Quasars with a high score are located near the peak of the distribution in red-side PCA coefficient space (i.e., "typical" quasars), while quasars with a low score are located in the outskirts of the distribution. The scores of the two $z > 7$ quasars are indicated by vertical lines. Despite the extreme nature of their C IV blueshifts relative to their systemic frames, the two $z > 7$ quasars are not extreme outliers of the score distribution, appearing at the 15.0 percentile (ULAS J1120+0641) and 1.5 percentile (ULAS J1342+0928).

One drawback of working in log space is that negative flux values are undefined—fortunately, the stacked auto-fit quasar continua we are using as our training set are essentially always positive because the true continua (in the absence of artifacts) should always be positive, and as mentioned above, we have explicitly thrown out the small fraction of quasars whose auto-fit continua fall below positive thresholds.

Our goal is to predict the shape of the Lyα region using the information contained in the rest of the spectrum. We adopt the projection procedure of Suzuki et al. (2005) and Pâris et al. (2011), wherein a linear mapping is constructed between the measured and predicted coefficients. Instead of fitting the red side and projecting to a combined red+blue continuum as performed by Suzuki et al. (2005) and Pâris et al. (2011), we keep the red- and blue-side spectra distinct, although in practice we find that this makes little difference. We first decompose the red- and blue-side spectra with the log-space PCA described above using the PCA implementation in the python SCIKIT-LEARN package (Pedregosa et al. 2011). We truncate the set of PCA basis spectra to the first 10 red-side and 6 blue-side basis spectra ($\boldsymbol{R}_i$ and $\boldsymbol{B}_i$ for the red and blue sides, respectively). The choice of the number of PCA basis spectra to keep is largely arbitrary—we chose 10 red-side basis spectra motivated by previous PCA analyses by Suzuki (2006) and Pâris et al. (2011), and then chose the number of blue-side basis spectra such that the error in the blue-side predictions (discussed later in Section 4) did not decrease with additional spectra. Tests with increased number of red-side and blue-side basis spectra (up to 15 and 10, respectively) showed very similar results, so our analysis is not particularly sensitive to the number of basis spectra.

In Figure 3 we show the red-side and blue-side mean of the log spectra (top row) and basis spectra (lower rows). Notably, the first red-side basis spectrum, $\boldsymbol{R}_1$, is a smooth curve that describes the broadband continuum variations between quasars. As mentioned above, if the variations between quasar continua were simply described by differences in spectral index (and uncorrelated with any other spectral features), the first basis spectrum should be linear in $\log\lambda$, and this is approximately the case.[12] The other red-side components encode correlations between the strengths of various broad emission lines and pseudo-continuum features, e.g., overlapping Fe II emission lines which blanket the spectrum. The blue-side components are more difficult to interpret directly, but the first two components appear to show either correlated ($\boldsymbol{B}_1$) or anti-correlated ($\boldsymbol{B}_2$) Lyα and N V line strengths.

To determine the relationship between the red-side and blue-side PCA coefficients (r_i and b_i, respectively), we now compute the coefficients for each individual (i.e., not median stacked) quasar spectrum in our training set of 12,764. We first fit for the red-side coefficients, r_i, via χ^2 minimization on the original (i.e., not auto-fit) spectra, after masking pixels which deviate

[12] In detail, there is a modest non-power-law curvature in $\boldsymbol{R}_1$. Interpreting this curvature is beyond the scope of this work, but we note that the form of the log-space decomposition is similar to those used to describe extinction curves.

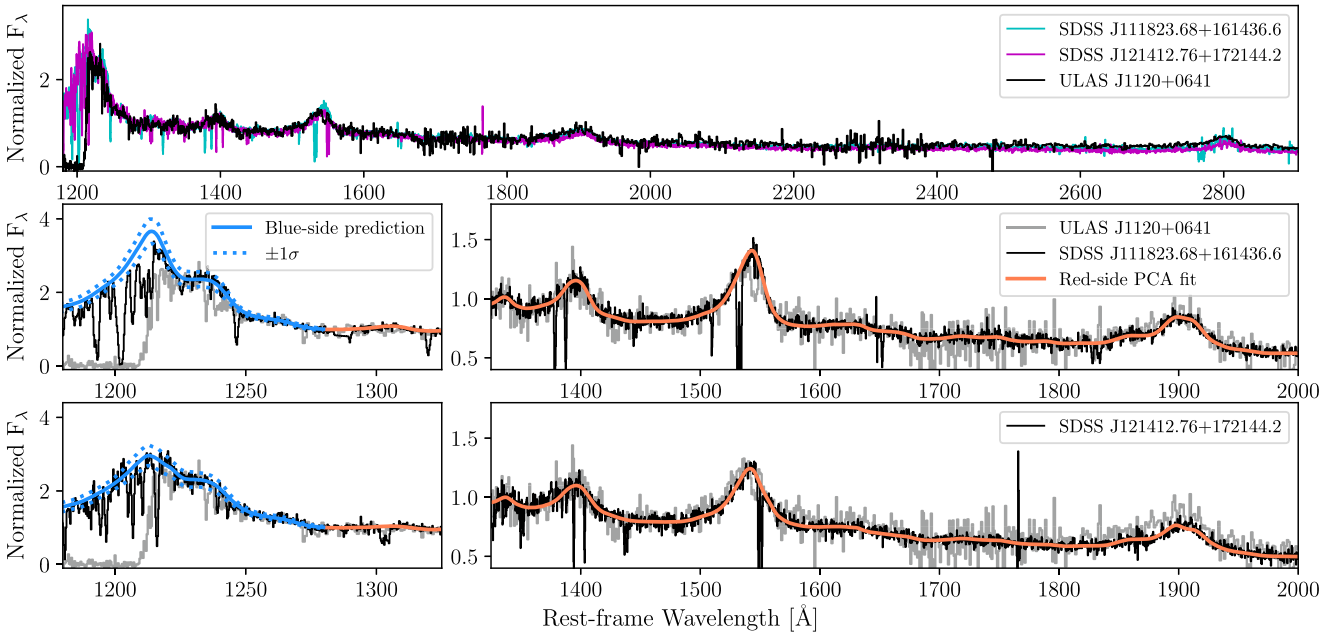

Figure 12. Top: comparison between the spectrum of ULAS J1120+0641 (black) and its two nearest neighbors in red-side PCA coefficient space: SDSS J111823.68+161436.6 (blue; $D_r = 1.24$) and SDSS J121412.76+172144.2 (magenta; $D_r = 1.58$). The wavelength axis is presented in the best-fit template redshift frame so that all three spectra shown have a consistently defined redshift. Lower panels: zoom in on the red-side fit (orange curve; right panels) and blue-side projection (blue curve; left panels) for the nearest-neighbor quasars. The dotted blue curves in the left panel show $\pm 1\sigma$ continuum prediction uncertainty.

more than 3σ from the auto-fit continua to remove metal absorption lines such as C IV and Mg II. Motivated by the large velocity shifts between the systemic-frame (given by, e.g., [C II] emission from the host galaxy) and the broad emission lines in $z \gtrsim 6.5$ quasars (e.g., Mazzucchelli et al. 2017), we simultaneously fit for a template redshift, i.e., we allow the effective redshift of the PCA basis spectra be a free parameter in the fit. This template redshift, $z_{\rm temp}$, can be consistently measured for any quasar spectrum with similar spectral coverage, allowing for direct comparison between our low-redshift training set and the high-redshift quasars we are interested in, and we ascribe no physical interpretation to its value. Then, working in the $z_{\rm temp}$ frame as defined by the red-side fit, we fit the blue-side coefficients, b_i, via χ^2 minimization on the auto-fit spectra instead of directly from the spectral pixels because the actual blue-side spectrum is guaranteed to be contaminated by Lyα forest absorption at $\lambda_{\rm rest} < \lambda_{{\rm Ly}\alpha}$.

The predictive power of the PCA decomposition lies in the correlations between r_i and b_i, shown in Figure 4, which show the joint distribution of best-fit coefficients of the training set spectra. In the left panel of Figure 5, the 2D histogram and corresponding blue contours show the relationship between r_2 and b_1, which appears to reflect a correlation between the strength of C IV (red-side) and Lyα (blue-side) emission line strengths (see Figure 3). Following Suzuki et al. (2005) and Pâris et al. (2011), we model these correlations between eigenvalues with a linear relationship, i.e., $b_i \approx \sum_{j=1}^{N_{\rm PCA,r}} r_j X_{ji}$, where $\boldsymbol{X}$ is the $N_{\rm PCA,r} \times N_{\rm PCA,b}$ projection matrix determined by solving the linear set of equations,

$$\boldsymbol{b} = \boldsymbol{r} \cdot \boldsymbol{X}, \tag{5}$$

where $\boldsymbol{b}$ ($N_{\rm spec} \times N_{\rm PCA,b}$) and $\boldsymbol{r}$ ($N_{\rm spec} \times N_{\rm PCA,r}$) are the sets of all best-fit blue-side and red-side coefficients from the training set, respectively. We solved for $\boldsymbol{X}$ using the least-squares solver in the python package `numpy` (van der Walt et al. 2011). The red contours in Figure 5 show the distribution of predicted b_1 (left) and b_4 (right) as a function of r_2 and r_6, respectively, after application of Equation (5) to the full set of r_i. While the projection matrix provides a close approximation to the relationship between b_1 and r_2, the relationship between b_4 and r_6 in the training set spectra has considerably more scatter in b_4 than the predicted values. This excess scatter may be related to either stochastic variations in the spectrum (e.g., the relationship between red-side and blue-side emission line properties is not exactly 1-to-1) or nonlinearities in the relationship between the blue-side and red-side components that are not captured by the linear model of Equation (5).

We show two examples of red-side PCA fits to quasars with S/N close to the median of our training set and their respective blue-side predictions in Figure 6. The information contained in the shapes and amplitude of the red-side emission lines is translated into the predicted blue-side spectrum through the mapping described by Equation (5), and for these quasars those predictions appear to be close to the true continuum. The quality of the red-side fits and blue-side predictions are only modestly affected by S/N, at least for the spectra selected with our S/N > 7 cut, and in Figure 7 we show example fits to quasars at the lower and upper ends of the distribution of S/N on the left (S/N ∼ 7) and right (S/N ∼ 25), respectively. To avoid only showing good examples, and in a sense of full disclosure, we show two examples of particularly bad blue-side predictions[13] in Figure 8. We quantify the general accuracy and precision of the blue-side predictions below.

4. Quantifying Uncertainties in the Continuum Predictions

There are several sources of uncertainty in the prediction of the blue-side continuum, including stochasticity in the relationship

[13] These poor predictions were discovered by inspecting spectra whose residuals at $\lambda_{\rm rest} \sim 1230$ Å were at the upper and lower ends of the distribution.

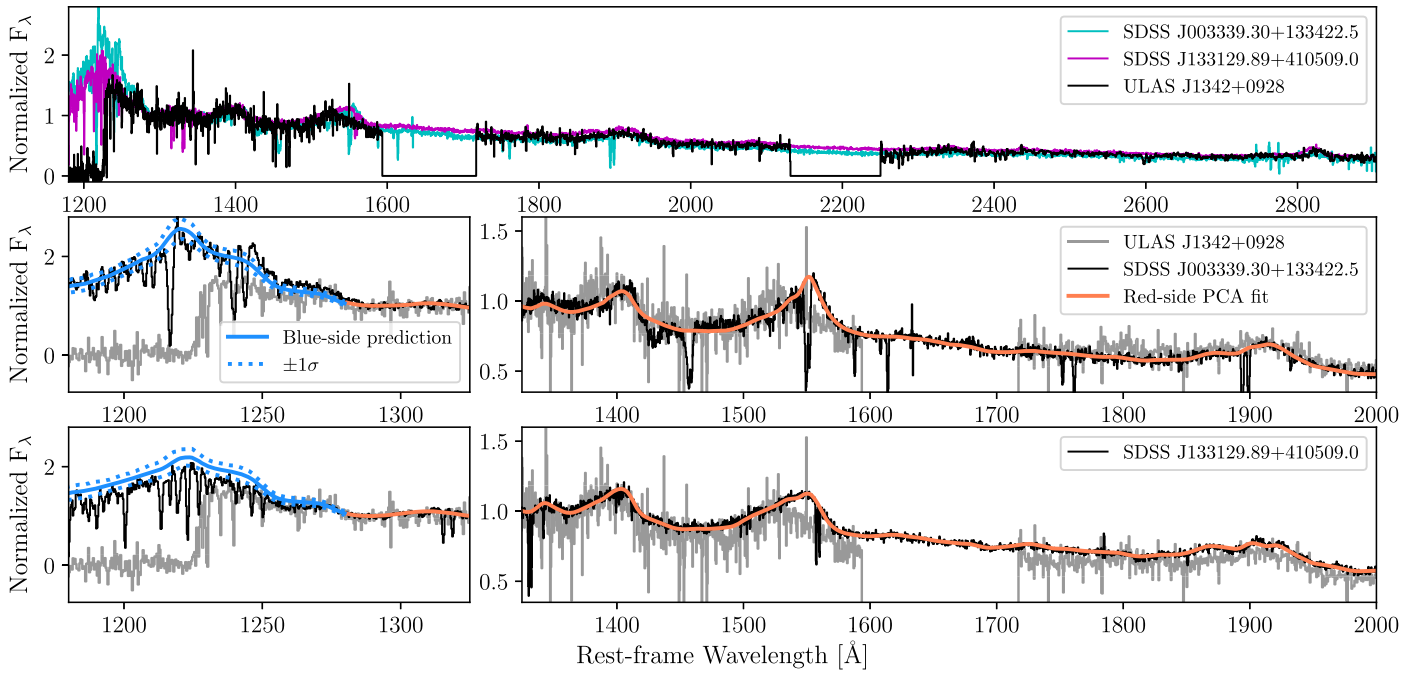

Figure 13. Similar to Figure 12, but for ULAS J1342+0928 and its two nearest neighbors: SDSS J003339.3+133422.5 (blue; $D_r = 2.30$) and SDSS J133129.89 +410509.0 (magenta; $D_r = 2.46$). The substantial difference between systemic and template redshift frames is evident from the onset of saturated Lyα absorption at $\lambda_{\rm rest} \sim 1225$ Å in ULAS J1342+0928. While globally the spectra look very similar, the C IV profiles of the nearest-neighbor quasars appear to differ somewhat from that of ULAS J1342+0928.

between red-side and blue-side features and the inability of the PCA model to exactly reproduce a given spectrum. We quantify the sum of these uncertainties by testing the full predictive procedure on every quasar in the training set, computing the relative continuum error of $\epsilon_C \equiv |F_{\rm pred} - F_{\rm true}|/F_{\rm pred}$, where $F_{\rm pred}$ is the predicted flux and $F_{\rm true}$ is the true flux. For the purposes of this analysis, we consider the auto-fit continuum of each quasar to be the true continuum, although this introduces an additional source of noise and a source of uncertainty in the actual true continuum that we do not attempt to quantify here. From simulations of mock Lyα forest absorption applied to noisy spectra at roughly half the resolution of the SDSS/BOSS spectra used here, Dall'Aglio et al. (2009) found that the auto-fit continua were biased low by a few percent in the Lyα forest at $z \sim 2$–2.5. In this work we ignore this bias because the Lyα forest bias only affects the spectrum blueward of Lyα, which is strongly absorbed in high-redshift quasars.

In the left panel of Figure 9, we show the mean and standard deviation of ϵ_C as thin and thick curves, respectively, as a function of rest-frame wavelength (where the rest frame of each quasar is defined by its best-fit template redshift, $z_{\rm temp}$) after three iterations of clipping individual pixels that deviate more than three standard deviations from the mean (resulting in ∼2% of all blue-side pixels masked). The 1σ error in the region most relevant to damping wing studies, $1210 < \lambda_{\rm rest} < 1250$ Å, is ∼6%–12%, comparable to the ∼9% error of the parametric method in Greig et al. (2017b)[14] but without requiring a strict prior over nearby redward flux that may bias against detecting extended damping wing absorption (i.e., high neutral fraction). We also note that Greig et al. (2017b) only considered quasars whose Lyα profiles were well modeled by their Gaussian fits, while in principle our method has more freedom to reproduce more complicated spectral morphologies.

Errors in the continuum prediction are strongly correlated across neighboring pixels, in part due to small correlated errors in the smoothed spline fit continua that we assume to be the true continua, but mostly due to smooth variations in the shape of broad emission lines and features of the underlying continuum reconstructed by the PCA model. We show the correlation matrix of ϵ_C in the right panel of Figure 9. The prediction uncertainty is strongly correlated on the scale of the spline fit (the roughly fixed width along the diagonal), and shows larger scale correlations due to variations in the strengths of broad N V ($\lambda_{\rm rest} \sim 1240$ Å) and Si II ($\lambda_{\rm rest} \sim 1260$ Å) emission lines. These strongly correlated continuum uncertainties, not limited to our method (Kramer & Haiman 2009), are a critical feature of quasar damping wing analyses that must be fully propagated when conducting parameter inference.

We also note that the red-side fits are not perfect, i.e., the PCA basis is unable to exactly reproduce the input spectra. Assuming again that the auto-fit continuum models represent the true continua, we show the 1σ relative error and mean bias of the red-side fits in the left panel of Figure 10. The error on the best-fit red-side PCA continuum is typically ∼3%, increasing smoothly above $\lambda_{\rm rest} \sim 2100$ Å to ∼5% at ∼2800 Å, close to the Mg II broad emission line. Interestingly, the typical red-side fit is biased by ∼1% across the entire wavelength range, except for small regions close to the peaks of the C IV and Mg II broad emission lines where the sign is reversed. This bias likely comes about because the actual pixels are used to fit the PCA model of the spectrum instead of the auto-fit continua, which may be biased slightly high due to differences in how outlier pixels are rejected. We also show the correlation matrix of the red-side fit errors in the right panel of Figure 10.

[14] Greig et al. (2017b) state that ∼90% of their predicted fluxes at $\lambda_{\rm rest} = 1220$ Å lie within 15% of the true continuum, corresponding to $\sim\pm1.64\sigma$ assuming Gaussian distributed errors.

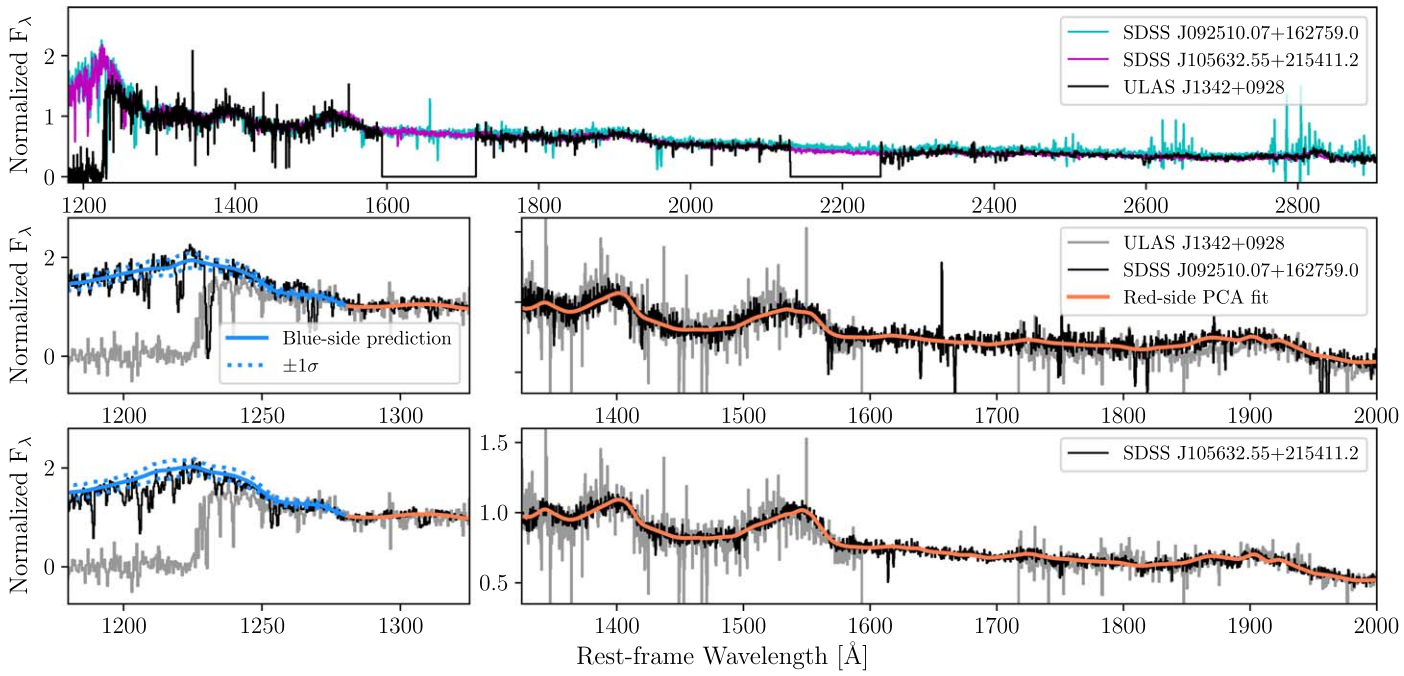

Figure 14. Similar to Figures 12 and 13, but for two other nearest-neighbor quasars to ULAS J1342+0928: SDSS J092510.07+162759.0 (blue; $D_r = 2.70$, 12th-nearest) and SDSS J105632.55+215411.2 (magenta; $D_r = 2.92$, 29th-nearest) which have been chosen by eye to have more similar C IV line profiles than the neighbors shown in Figure 13.

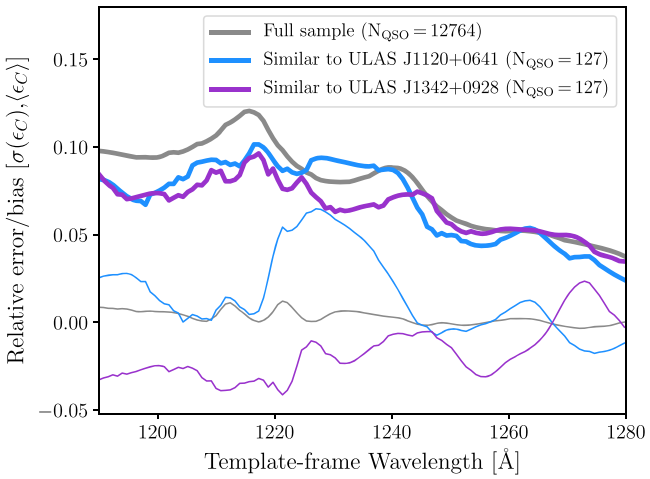

Figure 15. Relative uncertainty (1σ, thick curves) and mean bias (thin curves) for the 1% of quasars in the training set with red-side coefficients most similar to ULAS J1120+0641 (blue) and ULAS J1342+0928 (purple), compared to the full training set sample (gray).

5. Predicted Blue-side Continua of the Known $z > 7$ Quasars

We now demonstrate our continuum-fitting machinery on the two highest redshift quasars known: ULAS J1120+0641 ($z = 7.09$, Mortlock et al. 2011) and ULAS J1342+0928 ($z = 7.54$, Bañados et al. 2018). Modeling the intrinsic continuum of these quasars, and understanding the uncertainty in those models, is critical to constraining the neutral fraction of the IGM, so they provide a first test of the applicability of our continuum modeling procedure.

For our analysis of the two $z > 7$ quasars, we use the previously published spectra from their respective discovery papers. We use the combined VLT/FORS2 + Gemini/GNIRS spectrum of ULAS J1120+0641 published in Mortlock et al. (2011) and we use the combined *Magellan*/FIRE + Gemini/GNIRS spectrum of ULAS J1342+0928 published in Bañados et al. (2018). We fit the red-side continua of these quasars nearly identically to our fits of the training set, performing χ^2 minimization to fit both their red-side coefficients, r_i, and their template redshifts, $z_{\rm temp}$. As with the training set, we reject pixels that deviate from an automated spline fit of the continuum by $>3\sigma$ to reject strong metal absorption lines. In addition, for both spectra we mask spectral regions corresponding to regions of poor atmospheric transmission in the gaps between the *J*-/*H*-/*K*-bands, and for ULAS J1120+0641, due to the relatively low spectral resolution ($R \sim 500$) compared to the training set ($R \gtrsim 1200$), we manually mask regions of strong metal line absorption reported by Bosman et al. (2017) from their extremely deep VLT/X-Shooter spectrum. For ULAS J1120+0641 we find $z_{\rm temp} = 7.0834$, a very small blueshift of $\Delta v = 63\,{\rm km\,s^{-1}}$ from the systemic redshift ($z_{\rm sys} = 7.0851$ from host galaxy [C II] emission, Venemans et al. 2017b), while for ULAS J1342+0928 we find $z_{\rm temp} = 7.4438$, a blueshift of $\Delta v = 3422\,{\rm km\,s^{-1}}$ from the systemic redshift ($z_{\rm sys} = 7.5413$ from host galaxy [C II] emission, Venemans et al. 2017a).

Both of these quasars have peculiar broad emission line properties, most notably large blueshifts in the C IV line relative to Mg II: $\Delta v \sim 2800\,{\rm km\,s^{-1}}$ for ULAS J1120+0641, and $\Delta v \sim 6100\,{\rm km\,s^{-1}}$ for ULAS J1342+0928. We quantify the outlying nature of these quasars in the context of our PCA model by modeling the 10D probability distribution of the best-fit r_i from the training set as a mixture of multivariate Gaussians, i.e., a Gaussian mixture model (GMM), using the GaussianMixture package in scikit-learn. We chose the number of Gaussians, $N_{\rm Gauss} = 9$, to minimize the Bayesian information criterion (BIC; Schwarz 1978), defined by

$$\mathrm{BIC} = k \log N_{\rm spec} - 2 \log \hat{L}, \quad (6)$$

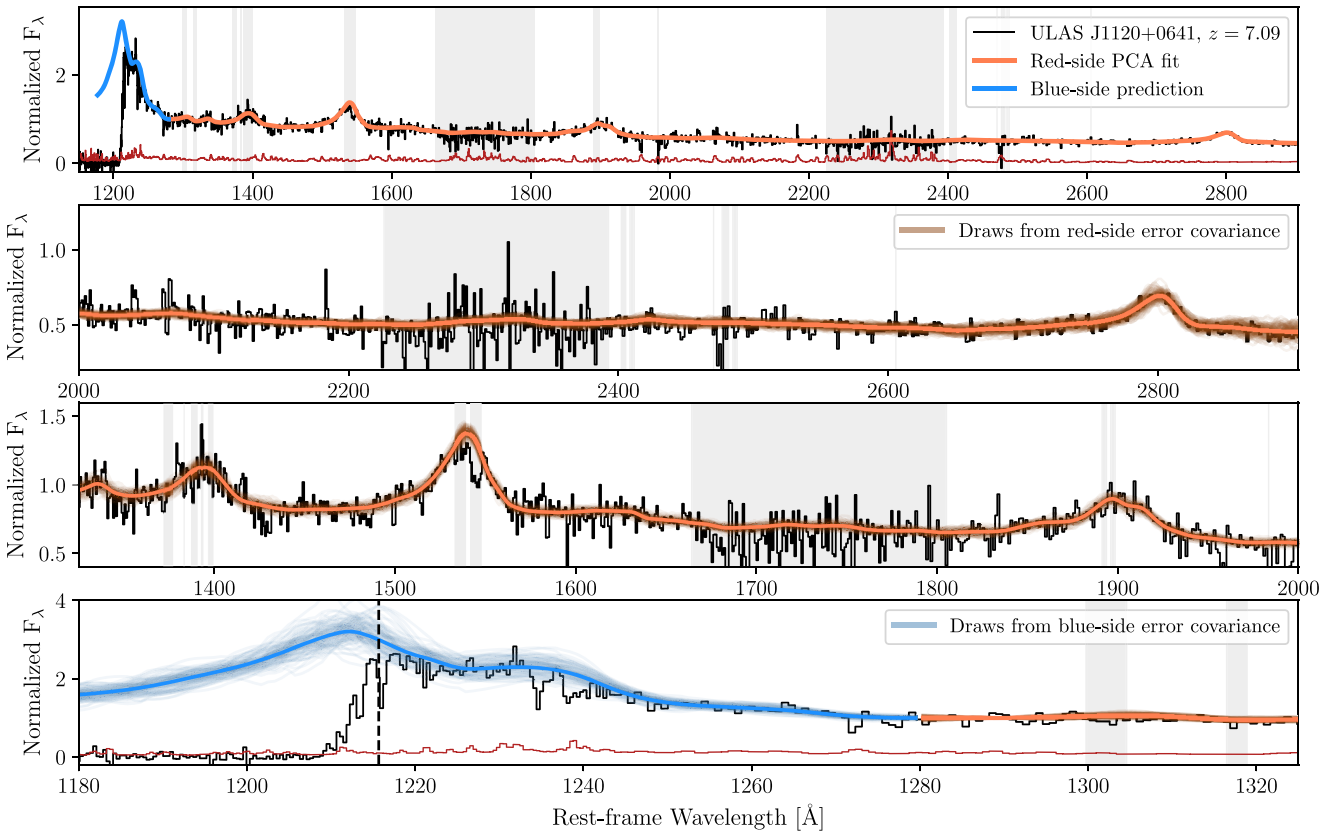

Figure 16. Top: FORS2+GNIRS spectrum of ULAS J1120+0641 (black) and its noise vector (red) from Mortlock et al. (2011). The red-side PCA fit and bias-corrected blue-side prediction are shown as the orange and blue curves, respectively. Gray shaded regions represent pixels that were masked out when performing the red-side fit. Middle: zoom in on the red-side fit of the strongest broad emission lines. The brown transparent curves show 100 draws from the covariant red-side fit error shown in Figure 10. Bottom: zoom in of the blue-side spectrum, where the vertical dashed line corresponds to rest-frame Lyα. The blue transparent curves show 100 draws from the covariant blue-side prediction error measured for 127 similar quasars in the training set.

where k is the number of model parameters of the GMM (i.e., means, amplitudes, and covariances of the individual Gaussians),

$$k = N_{\rm Gauss} \times \left(\frac{1}{2}N_{\rm PCA}^2 + \frac{3}{2}N_{\rm PCA} + 1\right), \quad (7)$$

$N_{\rm spec}$ is the number of spectra in the training set, and $\hat{L}$ is the maximum likelihood (i.e., the product of the GMM evaluated for each quasar) for the given number of Gaussians. The distribution of scores of each quasar in the training set, defined as the log of the GMM probability evaluated at their corresponding values of r_i, is shown in Figure 11. Quasars whose spectra have a higher score are located closer to the mean quasar spectrum, while lower scores represent outliers. The scores of ULAS J1120+0641 and ULAS J1342+0928 are shown by the vertical blue and purple lines corresponding to percentiles of ∼15.0% and 1.5% in the distribution of scores, respectively. Given the extreme broad emission line properties of these two quasars (Mortlock et al. 2011; Bosman & Becker 2015; Bañados et al. 2018), these percentiles may seem somewhat high—however, the GMM score is sensitive to more than just the particularly extreme features of the $z > 7$ quasar spectra (e.g., their large C IV blueshifts), illustrating that in other aspects the quasars are more representative of the bulk sample at low redshift (e.g., continuum slope and equivalent widths of broad emission lines).

Our ability to constrain the intrinsic blue-side continua of peculiar quasars like ULAS J1120+0641 and ULAS J1342+0928 can be determined by testing how well we can make predictions for similar quasars, i.e., with similar r_i, in the training set. We define a distance in r_i space by

$$D_r \equiv \sqrt{\sum_i^{N_{\rm PCA,r}} \left(\frac{\Delta r_i}{\sigma(r_i)}\right)^2}, \quad (8)$$

where i is the PCA component index, $N_{\rm PCA,r} = 10$, is the number of red-side PCA basis vectors, Δr_i is the difference between the best-fit r_i values, and $\sigma(r_i)$ is the standard deviation of best-fit r_i values in the training set sample. The median distance between randomly chosen quasars is $D_{r,\rm med} \sim 7.5$. We then identify the 1% of quasars ($N_{\rm QSO} = 127$) in the training set with the smallest D_r to the quasar whose continuum we are predicting, i.e., the set of nearest-neighbor quasar spectra in the training set. In Figures 12 and 13 we show the two nearest neighbors (i.e., the first and second smallest D_r) to ULAS J1120+0641 and ULAS J1342+0928, respectively, along with their respective red-side PCA fits and blue-side predicted continua. While the ULAS J1120+0641 neighbors have strikingly similar red-side broad emission line profiles, the ULAS J1342+0928 neighbors are less similar, reflecting the more sparse sampling of the distribution of quasar spectra. More qualitatively similar spectra exist in the set of nearest neighbors, however, and we show two examples in Figure 14,

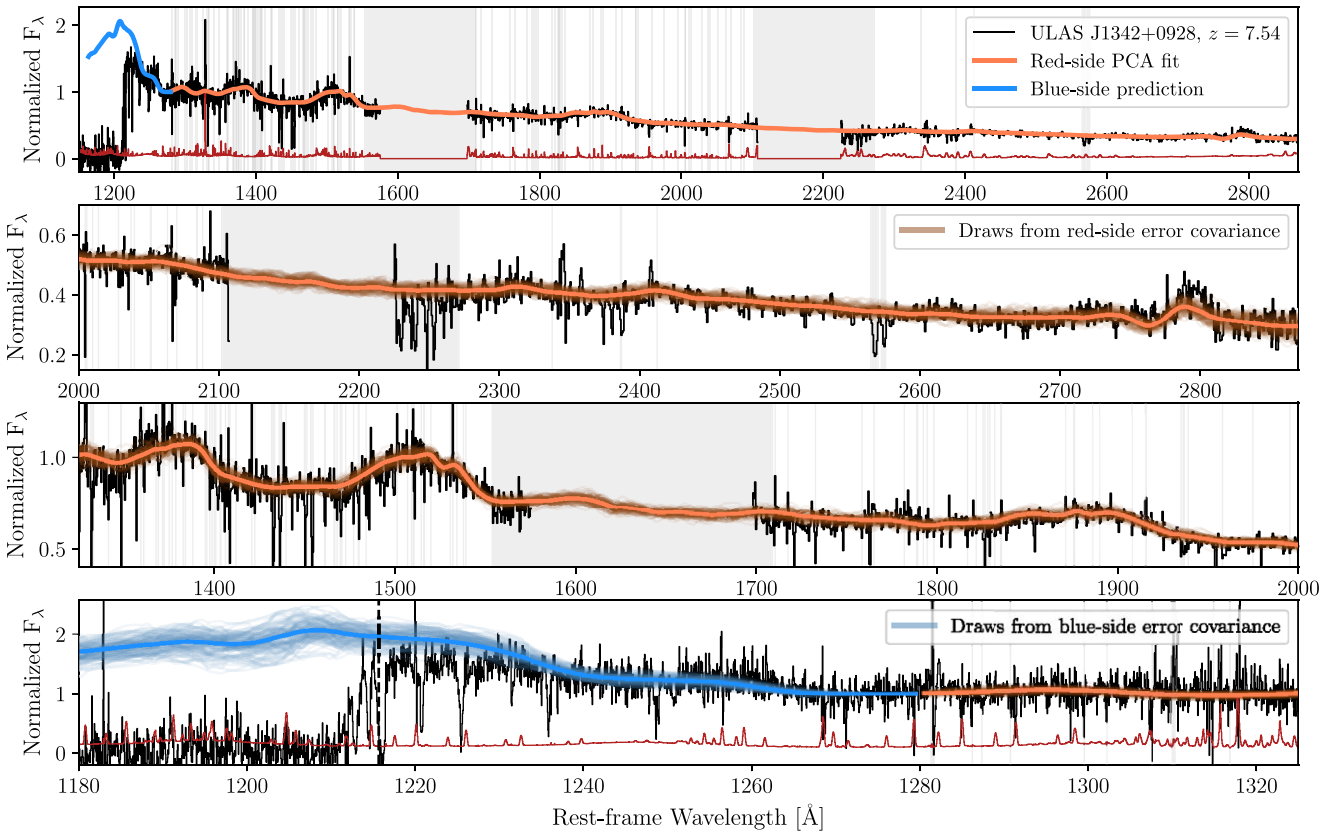

Figure 17. Similar to Figure 16, but for the FIRE+GNIRS spectrum of ULAS J1342+0928 from Bañados et al. (2018). The FIRE portion of the spectrum in the top two panels has been rebinned to match the pixel scale of the GNIRS data used in the *K*-band, while the bottom panel is shown at the original FIRE pixel scale to better highlight the smooth continuum regions close to Lyα between the narrow foreground absorption lines.

which were chosen by eye solely from a comparison of their red-side spectrum to that of ULAS J1342+0928.

The blue and purple curves in Figure 15 show the relative error and mean bias for the ULAS J1120+0641 and ULAS J1342+0928 nearest-neighbor samples, respectively, as a function of rest-frame ($z_{\rm temp}$-frame) wavelength. Quasars similar to ULAS J1342+0928 tend to have smaller continuum errors than average at $\lambda_{\rm rest} < 1245$ Å with a modest bias, while those similar to ULAS J1120+0641 tend to have error comparable to the full training set and a ∼5% bias at 1220 Å $\lesssim \lambda_{\rm rest} \lesssim$ 1235 Å. In Figures 12 through 14 we have corrected the blue-side predictions for their corresponding mean biases, and we similarly correct the predictions for the $z > 7$ quasars shown below.

To model the relative uncertainties between our blue-side predictions and the observed spectra, we approximate the continuum error distribution as a multivariate Gaussian distribution with mean and covariance set by the mean and covariance matrix of prediction errors ϵ_C measured from 127 nearest-neighbor quasars as described above. Our Gaussian approximation to the continuum errors appears to be a good one, and we show distributions of continuum error at representative rest-frame wavelengths in Appendix A. To generate plausible representations of the true continuum $F_{\rm sample}$, we draw samples of ϵ_C from this multivariate Gaussian and multiply the blue-side prediction by $1 + \epsilon_C$, i.e., $F_{\rm sample} = F_{\rm pred} \times (1 + \epsilon_C)$. This procedure can be used to draw Monte Carlo samples of the uncertainty in the continuum prediction for analyses of, e.g., the damping wing from the IGM.

In Figure 16 we show the red-side continuum fit (orange) and blue-side prediction (blue) for ULAS J1120+0641, where the latter has been corrected for the mean bias of the prediction for similar quasars (i.e., the thin blue curve in Figure 15). The pixel mask for the red-side fit is shown by the gray shaded regions. As shown in the middle panels, the PCA is able to fit the Si IV, C IV, C III], and Mg II broad emission lines very well, although most of the core of the C IV line has been masked out due to associated C IV absorbers which are unresolved in the GNIRS data. The blue-side prediction, shown more closely in the bottom panel, matches the blue-side continuum very closely for $\lambda_{\rm rest} > 1225$ Å, but at $\lambda_{\rm rest} \sim$ 1216–1225 Å there is evidence that the observed spectrum lies below the predicted continuum. This deficit suggests the presence of an IGM damping wing, as reported by Mortlock et al. (2011) and Greig et al. (2017a). However, from the covariant error draws $F_{\rm sample}$, shown as the transparent curves, it is clear that the uncertainty in our continuum model is of comparable amplitude to any putative damping wing signal, so any reionization constraint based solely on this damping wing signature will be weak at best.

In Figure 17 we show the red-side continuum fit and bias-corrected blue-side prediction for ULAS J1342+0928, analogous to Figure 16. The middle panels show that the PCA model is able to fit the red-side spectrum reasonably well despite the odd shape of the broad emission lines, although the fit to the Mg II line is quite poor.[15] Contrary to ULAS J1120+0641, and in agreement with the analysis in Bañados et al. (2018) based

[15] If we exclude the Mg II portion of the ULAS J1342+0928 spectrum from the red-side fit, we predict a very similar blue-side spectrum, so this does not appear to compromise the prediction procedure in general.

on a C IV-matched composite spectrum, the ULAS J1342 +0928 spectrum shows a significant, extended deficit relative to the blue-side prediction. This deficit is too large to explain with continuum error, and the deficit smoothly increases from $\lambda_{\rm rest} \sim 1250$ Å to $\lambda_{{\rm Ly}\alpha}$ in qualitative agreement with the expected profile of a Lyα damping wing from a substantially neutral IGM.

Finally, we note that the spectra of $z > 7$ quasars are typically taken with near-infrared echelle spectrographs, and so there may be additional systematic uncertainties in the red-side spectra (e.g., relative flux calibration uncertainty between echelle orders and telluric corrections) which are not present in our training set. Fully characterizing and quantifying these uncertainties is beyond the scope of this work, but in Appendix B we analyze two additional spectra of ULAS J1120+0641 taken with *Magellan*/FIRE (Simcoe et al. 2012) and VLT/X-Shooter (Barnett et al. 2017; Bosman et al. 2017). The continuum predictions vary between the three spectra at the $\sim$1–2σ level, but the differences between transmission spectra are somewhat smaller, hinting at the presence of differences between the intrinsic spectra at the different epochs at which the quasar was observed, although no substantial red-side variations were observed across different epochs of the X-Shooter data (G. Becker 2018, private communication).

In Davies et al. (2018) we determine constraints on the reionization epoch and quasar lifetimes from the continuum-normalized spectra of ULAS J1120+0641 and ULAS J1342 +0928, making use of the continuum error covariance matrices measured for their spectral nearest neighbors.

6. Conclusion

In this work, we have developed a PCA-based method for predicting the intrinsic quasar continuum at rest-frame wavelengths of $\lambda_{\rm rest} < 1280$ Å from the properties of the spectrum at 1280 Å $< \lambda_{\rm rest} < 2850$ Å. We exploited the large number of high-quality quasar spectra from SDSS/BOSS whose broad wavelength coverage enables us to build a continuous spectral model covering 1175 Å $< \lambda_{\rm rest} < 2900$ Å from a sample of 12,764 spectra with S/N > 7. After the initial processing of the spectra with adaptive, piecewise spline fits and subsequent nearest-neighbor stacking, we performed a log-space PCA decomposition of the training set truncated at 10 red-side and 6 blue-side basis spectra. We determined the best-fit values of these coefficients, and red-side template redshifts, for each quasar spectrum in the training set, and derived a projection matrix relating the red-side and blue-side coefficients. This projection matrix can then be used to predict the blue-side coefficients (and thus the blue-side spectrum) from a fit to the red-side coefficients (and template redshift) of an arbitrary quasar spectrum.

By testing our procedure on the training set, we found that we can predict the blue-side spectrum of an individual quasar to $\sim$6%–12% precision with very little mean bias ($\lesssim$1%), although prediction errors are strongly covariant across the entire blue-side spectrum. As a proof-of-concept test, we predicted the blue-side spectra of two $z > 7$ quasars thought to exhibit damping wings due to neutral hydrogen in the IGM: ULAS J1120+0641 (Mortlock et al. 2011) and ULAS J1342 +0928 (Bañados et al. 2018). These two quasars are known to possess outlying spectral features from the primary locus of quasar spectra at lower redshift, so we established that our method works similarly well on such outliers by testing the machinery on the 1% nearest-neighbor quasars in the training set to each $z > 7$ quasar. While ULAS J1120+0641 shows only modest evidence for an IGM damping wing, ULAS J1342 +0928 appears to be strongly absorbed redward of systemic-frame Lyα. In a subsequent paper we will constrain the neutral fraction of the IGM at $z > 7$ through statistical analysis of these two spectra.

A critical caveat of our analysis of $z > 7$ quasar spectra is that their near-infrared echelle spectra typically contain calibration artifacts that are not present in SDSS optical spectra. As shown in Appendix B, these systematic uncertainties may introduce additional scatter comparable to the blue-side prediction uncertainty measured from the training set. Future $z > 7$ quasar spectra taken by the *James Webb Space Telescope* should be free of these artifacts, and thus provide more definitive continuum predictions.

Our relatively unbiased method for quasar continuum prediction can be applied more broadly to quasar proximity zones at any redshift where direct measurement of the continuum is difficult, i.e., $z > 4$. Measurements of the quasar proximity effect using our predicted continua can in principle constrain the strength of the ionizing background (e.g., Dall'Aglio et al. 2008), the helium reionization history from the thermal proximity effect (Khrykin et al. 2017, J. F. Hennawi et al. 2018, in preparation), and timescales of quasar activity (Eilers et al. 2017). The predicted continua may also be useful for analyzing proximate absorption systems such as damped Lyα absorbers at $z \gtrsim 5$.

We would like to thank Daniel Stern for supporting the discovery of ULAS J1342+0928, and Monica Turner for consultation and support in the early efforts to model the intrinsic continuum of ULAS J1342+0928. We would also like to thank George Becker and Sarah Bosman for sharing the X-Shooter spectrum of ULAS J1120+0641.

E.P.F., B.P.V., and F. Walter acknowledge funding through the ERC grant "Cosmic Dawn."

Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is http://www.sdss3.org/.

SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofisica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, the Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University.

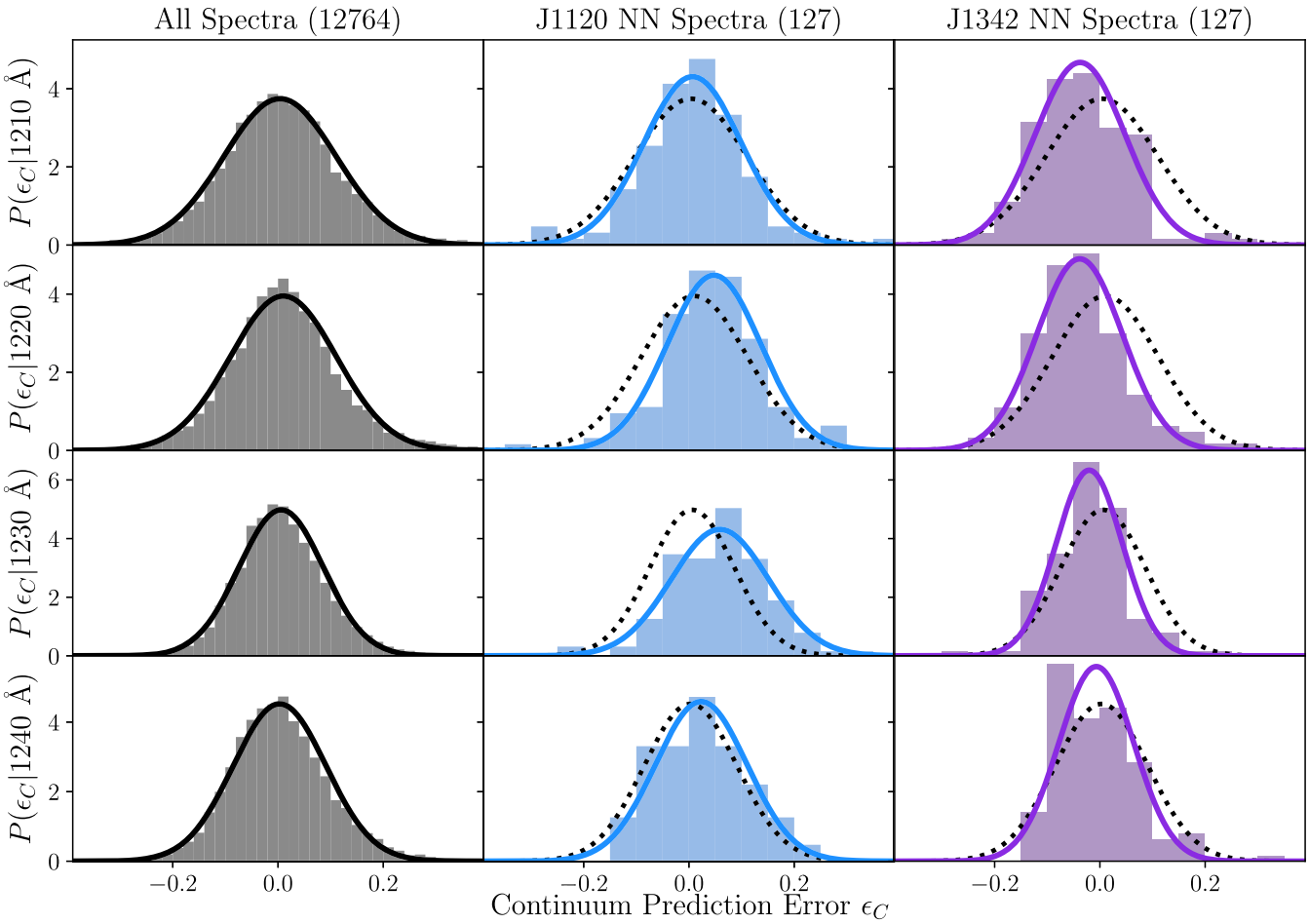

Figure 18. Distributions of continuum error measured for the full training set (left panels), nearest-neighbor spectra to ULAS J1120+0641 (middle panels), and nearest-neighbor spectra to ULAS J1342+0928 (right panels). From top to bottom, the error distributions are shown for $\lambda_{\rm rest} = 1210$, 1220, 1230, and 1240 Å.

Appendix A
Distribution of Continuum Prediction Errors

In Section 4 we approximated the distribution of continuum errors in the training set with a multivariate Gaussian, and in Section 5 we did the same for much smaller subsets of nearest-neighbor spectra to the $z > 7$ quasars. In Figure 18 we show representative distributions of continuum error in the training set (left panels) and the two nearest-neighbor samples (middle and right panels) at $\lambda_{\rm rest} = 1210$, 1220, 1230, and 1240 Å from top to bottom. The distributions (histograms) are well approximated by the Gaussian approximations (solid curves), and in each of the nearest-neighbor panels we show the distribution for the entire training set as a dotted curve, highlighting the significant differences in the error properties (i.e., the mean and variance) of the different subsets of quasars.

Appendix B
Calibration Systematics in $z > 7$ Quasar Spectra

The broadband spectra of $z > 7$ quasars are typically taken with near-infrared echelle spectrographs, potentially introducing significant order-by-order flux calibration and telluric correction uncertainties. While our log-PCA method should correct for any single power law tilt across the entire spectrum, order specific calibration systematics will not be represented by the SDSS-based training set that we use to quantify continuum prediction uncertainties.

In an attempt to qualitatively estimate the effect of these systematic uncertainties, we have applied our methodology to two additional independent spectra of ULAS J1120+0641 taken with *Magellan*/FIRE (Simcoe et al. 2012) and VLT/X-Shooter (Barnett et al. 2017; Bosman et al. 2017). We show the three spectra in Figure 19, normalized to the flux at $\lambda_{\rm rest} = 1290$ Å, and we have smoothed the FIRE and X-Shooter spectra with a three-pixel median filter for ease of comparison by eye. We can see subtle differences between the three spectra, most obviously a slight power law tilt in the FIRE spectrum relative to the other two. We fit only $\lambda_{\rm rest} < 2810$ Å due to artifacts in the *K*-band in the FIRE and X-Shooter spectra, and we refit the GNIRS data over the same restricted wavelength range for consistency. In Figure 20 we show the red-side fits to each spectrum, making the (PCA fit) differences more visible. The relative differences between the red-side fits are shown in the left panel of Figure 21. Interestingly, the largest relative differences between the red-side fits are evidently inside of broad emission lines, hinting at a possible physical (rather than instrumental) origin to the differences between spectra, although we note that broad emission lines dominate the red-side PCA eigenspectra (Figure 3), so they may be more sensitive to small-scale artifacts in the data. However, especially given the time dilation due to the high redshift of ULAS J1120+0641, the FIRE

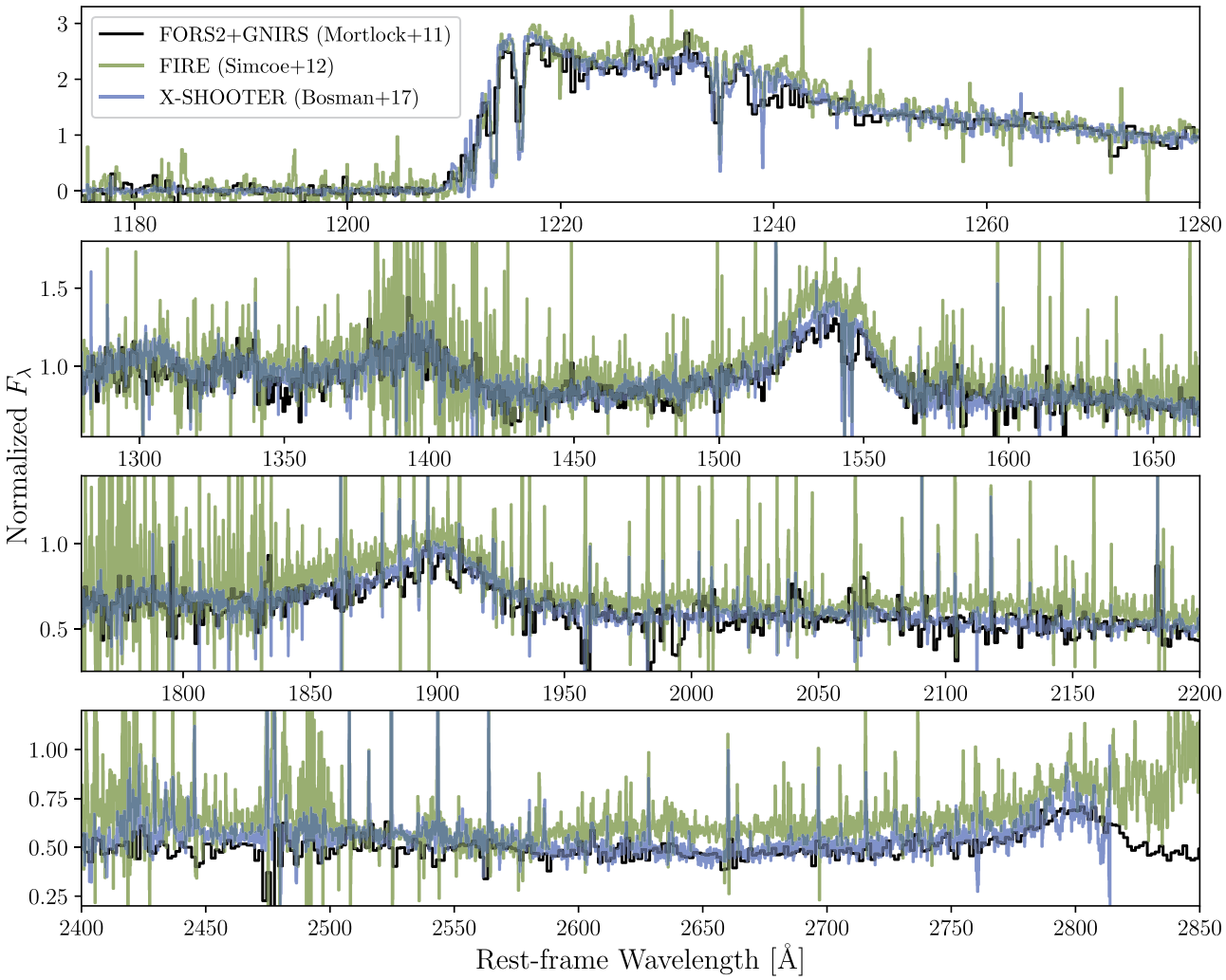

Figure 19. Comparison between the three independent spectra of ULAS J1120+0641. Our fiducial FORS2+GNIRS spectrum is shown in black, while the *Magellan*/FIRE (Simcoe et al. 2012) and VLT/X-Shooter (Barnett et al. 2017; Bosman et al. 2017) spectra are shown in green and blue, respectively. The FIRE and X-Shooter spectra have been median filtered over a three-pixel window for ease of comparison. Rest-frame wavelengths in regions with strong telluric absorption are not shown.

spectrum differs enough from the GNIRS and X-Shooter spectra that calibration artifacts seem to be a more likely explanation.

The resulting blue-side continuum predictions are shown in the right panel of Figure 21. The FORS2+GNIRS prediction differs from the fiducial prediction in the main text due to the difference in the fitted wavelength range, but the differences are $\lesssim 1\sigma$. The predictions for the FIRE and X-Shooter spectra differ from the FORS2+GNIRS continuum prediction by $\sim$1–2σ, which at first glance seems to suggest that we may have underestimated the continuum uncertainties by a factor of $\sim$2. In Figure 22 we show that the differences between the continuum-normalized spectra are smaller—in fact, the difference between the FORS2+GNIRS and FIRE spectra largely disappears at $\lambda_{\rm rest} < 1230$ Å. This agreement suggests a possible physical origin for the disagreement between the continuum models, i.e., joint variability between broad emission lines across the spectrum, but no such variability was seen across different years of the X-Shooter data (G. Becker 2018, private communication). Additionally, there are conspicuous regions where the transmission spectra differ substantially between spectra, in particular at $\lambda_{\rm rest} \sim 1218$ Å and $\sim$1242 Å where the differences (roughly 10%–20%) are comparable to the continuum prediction uncertainty measured from the training set. Precisely quantifying the additional uncertainty due to calibration artifacts is beyond the scope of this work, but we note that all three continuum models suggest the presence of a damping wing from neutral hydrogen along this sightline, and similarly consistent with the best-fit absorption model in Davies et al. (2018).

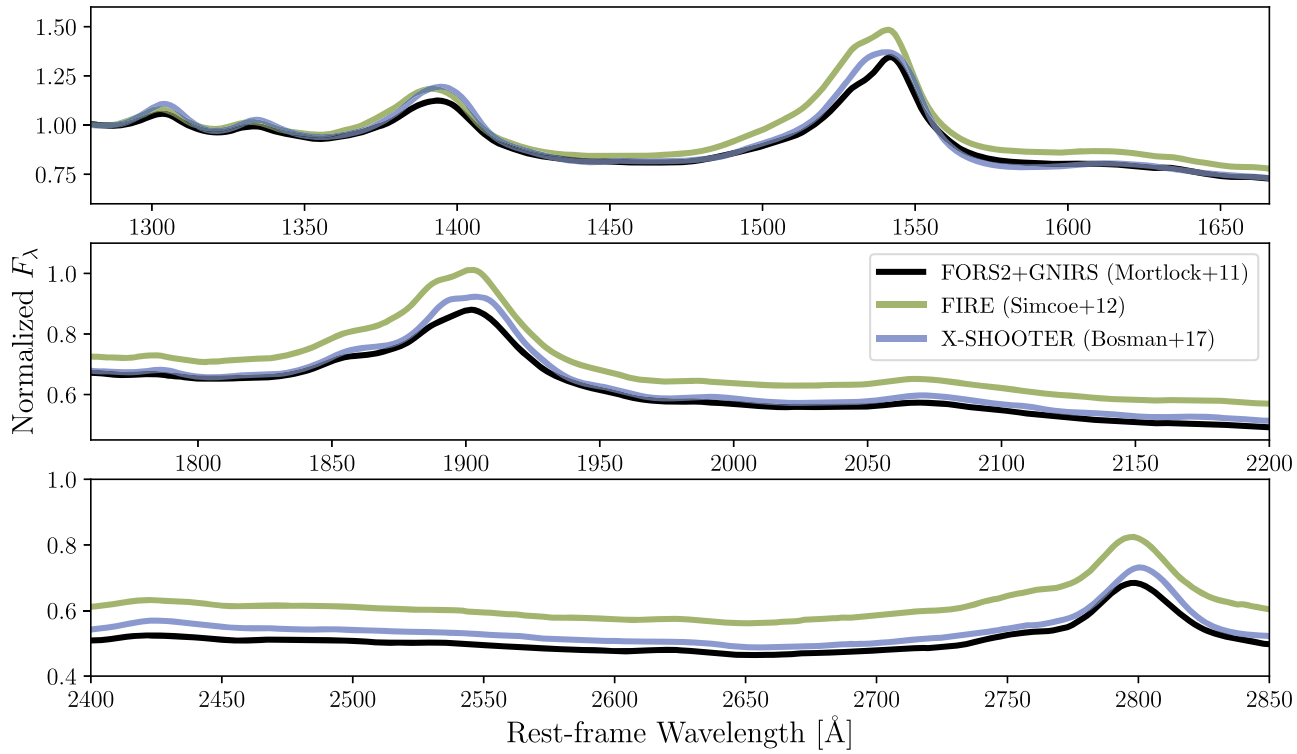

Figure 20. Similar to Figure 19, but now showing the red-side PCA fits of the ULAS J1120+0641 spectra.

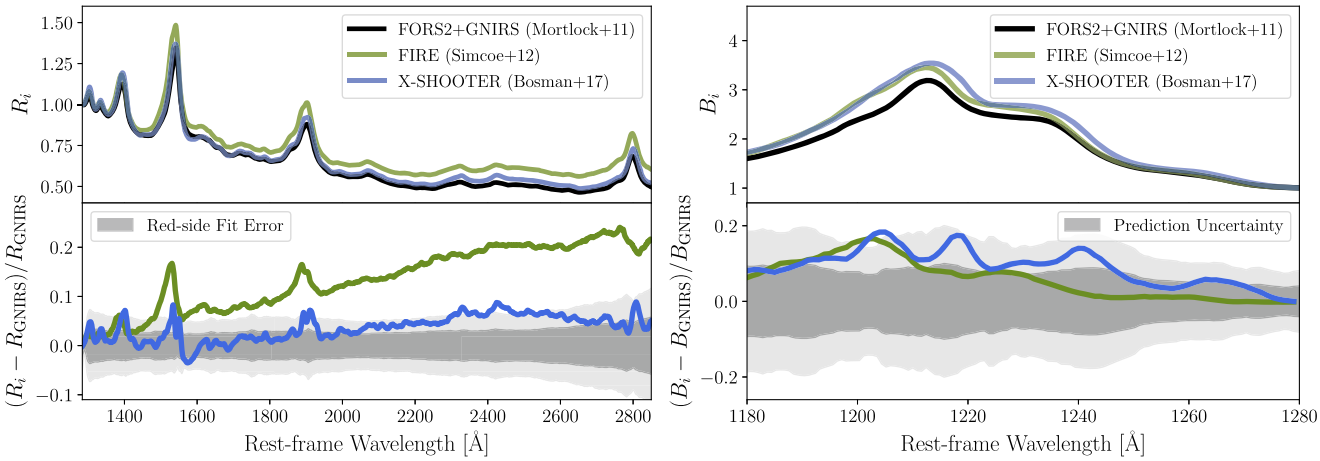

Figure 21. Left: red-side PCA fits of the ULAS J1120+0641 spectra (top panel) and the relative difference between the FORS2+GNIRS red-side fit and the FIRE (green) and X-Shooter (blue) fits (bottom panel). The 1σ and 2σ error in the red-side fits of the training set are shown by the dark and light shaded regions, respectively. Right: same as the left panels, but for the blue-side continuum predictions.

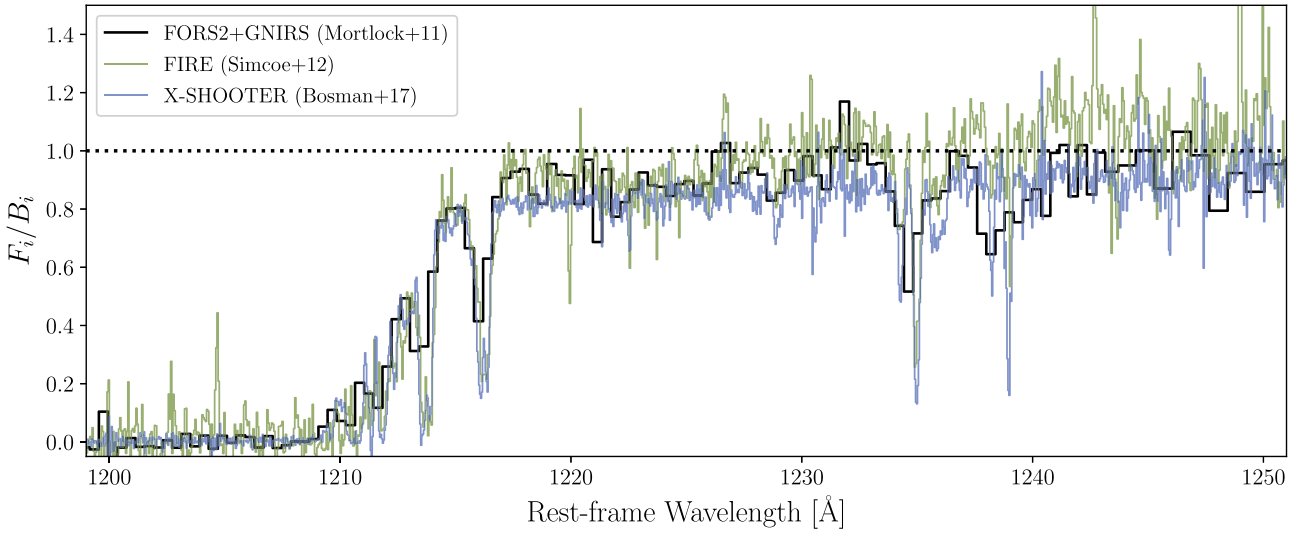

Figure 22. Transmission spectra of ULAS J1120+0641 from the FORS2+GNIRS (black), FIRE (green), and X-Shooter (blue) analyses.

ORCID iDs

Frederick B. Davies https://orcid.org/0000-0003-0821-3644
Joseph F. Hennawi https://orcid.org/0000-0002-7054-4332
Eduardo Bañados https://orcid.org/0000-0002-2931-7824
Robert A. Simcoe https://orcid.org/0000-0003-3769-9559
Roberto Decarli https://orcid.org/0000-0002-2662-8803
Xiaohui Fan https://orcid.org/0000-0003-3310-0131
Emanuele P. Farina https://orcid.org/0000-0002-6822-2254
Chiara Mazzucchelli https://orcid.org/0000-0002-5941-5214
Hans-Walter Rix https://orcid.org/0000-0003-4996-9069
Bram P. Venemans https://orcid.org/0000-0001-9024-8322
Fabian Walter https://orcid.org/0000-0003-4793-7880
Feige Wang https://orcid.org/0000-0002-7633-431X
Jinyi Yang https://orcid.org/0000-0001-5287-4242

References

Alam, S., Albareti, F. D., Allende Prieto, C., et al. 2015, ApJS, 219, 12
Bañados, E., Venemans, B. P., Mazzucchelli, C., et al. 2018, Natur, 553, 473
Barnett, R., Warren, S. J., Becker, G. D., et al. 2017, A&A, 601, A16
Boroson, T. A., & Green, R. F. 1992, ApJS, 80, 109
Bosman, S. E. I., & Becker, G. D. 2015, MNRAS, 452, 1105
Bosman, S. E. I., Becker, G. D., Haehnelt, M. G., et al. 2017, MNRAS, 470, 1919
Carswell, R. F., Whelan, J. A. J., Smith, M. G., Boksenberg, A., & Tytler, D. 1982, MNRAS, 198, 91
Dall'Aglio, A., Wisotzki, L., & Worseck, G. 2008, A&A, 491, 465
Dall'Aglio, A., Wisotzki, L., & Worseck, G. 2009, arXiv:0906.1484
Davies, F. B., Hennawi, J. F., Bañados, E., et al. 2018, ApJ, 864, 142
Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. 2013, AJ, 145, 10
Eilers, A.-C., Davies, F. B., Hennawi, J. F., et al. 2017, ApJ, 840, 24
Eisenstein, D. J., Weinberg, D. H., Agol, E., et al. 2011, AJ, 142, 72
Francis, P. J., Hewett, P. C., Foltz, C. B., & Chaffee, F. H. 1992, ApJ, 398, 476
Greig, B., Mesinger, A., Haiman, Z., & Simcoe, R. A. 2017a, MNRAS, 466, 4239
Greig, B., Mesinger, A., McGreer, I. D., Gallerani, S., & Haiman, Z. 2017b, MNRAS, 466, 1814
Harris, D. W., Jensen, T. W., Suzuki, N., et al. 2016, AJ, 151, 155
Khrykin, I. S., Hennawi, J. F., & McQuinn, M. 2017, ApJ, 838, 96
Kramer, R. H., & Haiman, Z. 2009, MNRAS, 400, 1493
Lee, K.-G., Suzuki, N., & Spergel, D. N. 2012, AJ, 143, 51
Mazzucchelli, C., Bañados, E., Venemans, B. P., et al. 2017, ApJ, 849, 91
Miralda-Escudé, J. 1998, ApJ, 501, 15
Mortlock, D. J., Warren, S. J., Venemans, B. P., et al. 2011, Natur, 474, 616
Pâris, I., Petitjean, P., Rollinde, E., et al. 2011, A&A, 530, A50
Pâris, I., Petitjean, P., Ross, N. P., et al. 2017, A&A, 597, A79
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, JMLR, 12, 2825
Prochaska, J. X. 2017, A&C, 19, 27
Richards, G. T., Kruczek, N. E., Gallagher, S. C., et al. 2011, AJ, 141, 167
Schwarz, G. 1978, AnSta, 6, 461
Shang, Z., Wills, B. J., Wills, D., & Brotherton, M. S. 2007, AJ, 134, 294
Simcoe, R. A., Sullivan, P. W., Cooksey, K. L., et al. 2012, Natur, 492, 79
Smee, S. A., Gunn, J. E., Uomoto, A., et al. 2013, AJ, 146, 32
Suzuki, N. 2006, ApJS, 163, 110
Suzuki, N., Tytler, D., Kirkman, D., O'Meara, J. M., & Lubin, D. 2005, ApJ, 618, 592
van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, CSE, 13, 22
Venemans, B. P., Walter, F., Decarli, R., et al. 2017a, ApJL, 851, L8
Venemans, B. P., Walter, F., Decarli, R., et al. 2017b, ApJ, 837, 146
Yip, C. W., Connolly, A. J., Vanden Berk, D. E., et al. 2004, AJ, 128, 2603
Young, P. J., Sargent, W. L. W., Boksenberg, A., Carswell, R. F., & Whelan, J. A. J. 1979, ApJ, 229, 891