PREDICTING QSO CONTINUA IN THE Lyα FOREST
NAO SUZUKI, DAVID TYTLER, DAVID KIRKMAN, JOHN M. O’MEARA, AND DAN LUBIN
Received 2003 June 27; accepted 2004 September 15

ABSTRACT
We present a method to make predictions with sets of correlated data values, in this case QSO flux spectra. We predict the continuum in the Lyα forest of a QSO, from 1020 to 1216 Å, using the spectrum of that QSO from 1216 to 1600 Å. We find correlations between the unabsorbed flux in these two wavelength regions in the Hubble Space Telescope (HST) spectra of 50 QSOs. We use principal component analysis to summarize the variety of these spectra, relate the weights of the principal components for 1020–1600 Å to the weights for 1216–1600 Å, and apply this relation to make predictions. We test the method on the HST spectra and find an average absolute flux error of 9%, with a range of 3%–30%, where individual predictions are systematically too low or too high. We mention several ways in which the predictions might be improved.

Subject headings: intergalactic medium — methods: data analysis — methods: statistical — quasars: absorption lines — quasars: emission lines — techniques: spectroscopic

Online material: color figures

1. INTRODUCTION
We would like to have an accurate and objective way to find the continuum level in QSO spectra, because these levels are required to measure the amount of absorption. The uncertainty in the continuum level is often one of the largest uncertainties in the studies of the intergalactic medium (IGM; e.g., Croft et al. 2002), and the precise continuum shape is required for precision measurement of absorption lines, such as the measurement of D/H (Kirkman et al. 2003; Suzuki et al. 2003).

Standard methods of estimating the continua in the Lyα forest region of QSO spectra are frequently unsatisfactory. For redshifts 2 < z < 4, the standard way to find a continuum level is to fit a smooth curve over the peaks of the flux in the Lyα forest. This method works well, giving less than 2% error in the continuum level in high-resolution spectra (e.g., 8 km s⁻¹ FWHM) with high S/N (e.g., 100 per 2 km s⁻¹; Kirkman et al. 2003). However, the method fails when there are few pixels that we can clearly identify as unabsorbed continuum, and this lack of continuum information is common in low-S/N spectra, in low-resolution spectra in which lines blend together, and at higher redshifts at which the Lyα forest absorbs more than a few percent at all wavelengths. In fact, for redshifts z > 6 the complete Gunn-Peterson trough (Becker et al. 2001; Djorgovski et al. 2001; Fan et al. 2001) makes it impossible to directly measure the continuum. Instead, at high redshifts, especially z > 4, it is common to approximate the continuum in the Lyα forest with a power-law extrapolation from wavelengths larger than Lyα (Telfer et al. 2002; White et al. 2003). But the continuum in the Lyα forest is not a power law, because of the wings of the Lyβ–O VI emission line, and especially the Lyα line, extend far into the Lyα forest, and there exist weak emission lines, especially near 1073 and 1123 Å (Zheng et al. 1997; Vanden Berk et al. 2001; Bernardi et al. 2003).

We may be able to predict the unabsorbed flux in the Lyα forest if it is correlated with the unabsorbed flux at other wavelengths. Here, the unabsorbed flux includes both continuum and emission lines but not the random intervening absorption. We measure correlations in a set of QSO spectra (the training set) that cover both 1020 Å ≤ λ ≤ 1215 Å (the blue side) and 1216 Å ≤ λ ≤ 1600 Å (the red side). We use the red-side spectra of individual QSOs to make predictions of their blue sides.

We use principal component analysis (PCA) to summarize the information in the QSO spectra. PCA seeks to reduce the dimensionality of large data sets and is widely used in astronomy (Whitney 1983; Kanbur et al. 2002; Efstatiiou 2002). Francis et al. (1992) applied PCA to the Large Bright Quasar Survey (LBQS; Hewett et al. 1995, 2001) optical spectra of QSOs to give an objective classification scheme, and they showed that any normalized QSO spectrum, q₁(λ), is well represented by a reconstructed spectrum, r₁, m(λ), that is a weighted sum of m principal components:

\[ q₁(λ) \sim r₁, m(λ) = \mu(λ) + \sum_{j=1}^{m} c_j \xi_j(λ), \]

where i refers to a QSO, \( \mu(λ) \) is the mean of many QSO spectra, \( \xi_j(λ) \) is the jth principal component, and \( c_j \) is its weight. Instead of classifying QSO spectra, we use PCA to make predictions.

In § 2 we describe Hubble Space Telescope (HST) spectra that we use for the training set. We show the correlations in the QSO spectra and results of the PCA in § 3. In § 4 we show how we make predictions, and we discuss their accuracy.

2. QSO SPECTRA AND THEIR CORRECTION
For the training set we use the UV spectra of low-redshift (z < 1) QSOs because they have little absorption and we can clearly see their continuum levels. Here we describe these spectra, the criteria that we use to select them, and the corrections that we make.

We use a subset of the 334 high-resolution (HST) Faint Object Spectrograph spectra collected and calibrated by Bechtold et al. (2002). This sample includes all the high-resolution QSO spectra from the HST QSO Absorption Line Key Project (Bahcall et al. 1993, 1996; Jannuzi et al. 1998). The gratings chosen are...
G130H, G190H, and G270H, and their spectral resolution is $R \sim 1300$. Bechtold et al. (2002) identified both IGM and interstellar medium lines in a uniform manner, and they applied Galactic extinction corrections using the Galactic reddening map of Burstein & Heiles (1982) and the Milky Way reddening curve of Cardelli et al. (1989).

We select QSO spectra by wavelength coverage and S/N, and we remove a few QSOs with peculiar spectra. We reject QSOs that do not have complete coverage from 1020 to 1600 Å. This range covers from the Ly$\beta$+O vi emission-line blend to the C iv emission line. A larger range would help reveal the shapes of the QSO continua, but we would then have fewer QSOs in the sample.

We reject QSOs that do not have an average S/N $> 10$ per binned pixel (0.5 Å in the rest frame) from 1050 to 1170 Å. We are interested in the intrinsic variation of the QSO spectra against the mean spectrum. Photon noise adds variation, masks the intrinsic variations, and alters the primary principal components. Before we removed the low-S/N spectra, we found that some principal components were largely reproducing the photon noise of the spectra with unusually low S/N.

We remove QSOs with broad absorption lines and damped Ly$\alpha$ systems because we are unsure where to place their continua. We also remove Q0219+4248 and Q0906+4305, whose emission-line features are extremely weak. These removals make our sample not representative of all QSO spectra.

We end up using the spectra of 50 QSOs, which we list in Table 1. The mean redshift is 0.58, with a standard deviation of 0.27 and a range from 0.14 to 1.04. The average S/N is 19.5 per 0.5 Å binned pixel.

We represent the spectra of all 50 QSOs by fitted smooth curves that reduce the effects of photon noise and interpolate over the absorption lines. To find the smooth curves, we mask the absorption lines that Bechtold et al. (2002) identified in both the blue and red sides. Then, for every 50 Å interval, with 20 Å overlaps, we fit Chebyshev polynomials and choose the order of the polynomials so that the reduced $\chi^2$ becomes close to unity. The order is about 4–6 if no strong emission lines lie in that interval. In intervals that include strong emission lines such as Ly$\alpha$ and C iv, the order becomes 30–40. For a few QSOs we made further adjustments using the software described in Kirkman et al. (2003) by fitting a smooth $b$-spline curve to the QSO continua.

In Table 1 we give emission redshifts that we measure from the peaks of the Ly$\alpha$ emission lines. In the rest frame the emission-line peaks align, and the asymmetric profiles become a part of the variance in the set of spectra. If we do not use the peak of the Ly$\alpha$ line but instead cross-correlate with a template of known redshift, we find that we need extra principal components to reconstruct the emission lines. Once we obtain the redshifts, we shift the spectra to the rest frame and rebin them into 0.5 Å pixels. We then have 1161 pixels of flux data values per QSO in the range 1020–1600 Å.

Since we are interested in the relative spectrum shape, we throw away absolute flux information. We find the average flux in 21 pixels around 1280 Å, and we normalize all spectra to unit flux at these wavelengths, far from any strong emission lines.

### 3. PRINCIPAL COMPONENT ANALYSIS OF QSO SPECTRA

We calculate the correlation of the fluxes at different wavelengths to see how different parts of the typical QSO spectrum are related. In Figure 1 we see the 1161×1161 correlation matrix $R$ with elements

$$R(\lambda_m, \lambda_n) = \frac{1}{N-1} \sum_{i=1}^{N} \frac{[q_i(\lambda_m) - \mu(\lambda_m)][q_i(\lambda_n) - \mu(\lambda_n)]}{\sigma(\lambda_m)\sigma(\lambda_n)},$$

where $q_i(\lambda)$ is the continuum-fitted and normalized spectrum for the $i$th QSO, $N$ is the total number of QSO spectra, and $\sigma_m$ and $\sigma_n$ are the standard deviations of the flux in the $m$th and $n$th wavelength bins, $\lambda_m$ and $\lambda_n$, respectively.

We find moderate correlation, about 0.2–0.6, between the red and blue continua. The correlation between the emission lines, 0.8, is much stronger, and hence we expect that the emission lines in the red side will give good predictions for those in the blue side.

We can calculate the covariance matrix $V$ for the 50 QSOs as

$$V(\lambda_m, \lambda_n) = \frac{1}{N-1} \sum_{i=1}^{N} [q_i(\lambda_m) - \mu(\lambda_m)][q_i(\lambda_n) - \mu(\lambda_n)].$$

In Figure 2 we see that the covariance is relatively small in the continuum but large in the emission lines, meaning that the emission lines vary a lot from QSO to QSO. The peaks near 1071, 1123, and 1176 Å probably correspond to the weak emission lines in the Ly$\alpha$ forest that Bernardi et al. (2003) discuss.

We can find the principal components by decomposing the covariance matrix $V$ into the product of the orthonormal matrix $P$, which is composed of eigenvectors, and the diagonal matrix $\Lambda$, containing the eigenvalues:

$$V = P^{-1} \Lambda P.$$

We call the eigenvectors (the columns of the matrix $P$) the principal components. The principal components are ordered according to the amount of the variance in the training set that they can accommodate, such that the first principal component is the eigenvector that has the largest eigenvalue.

Let us quantitatively assess how well we can reconstruct the QSO spectra using a certain number of components. We find the weight $c_j$ of the $j$th principal component of a QSO spectrum $q_i(\lambda)$ from

$$c_j = \int_{1020\ \text{Å}}^{1600\ \text{Å}} [q_i(\lambda) - \mu(\lambda)]\xi_j(\lambda) \, d\lambda.$$  

When we use the first $m$ components, we get the reconstructed spectrum $r_{i,m}$. The $\xi_j(\lambda)$ look similar to QSO spectra but with more structure at the wavelengths of the emission lines. Examples of principal components and reconstructions, which are very similar to ours, are given by Francis et al. (1992).

We now introduce the accumulated residual variance fraction $\delta E_{i,m}$:

$$\delta E_{i,m} = \frac{\int_{1020\ \text{Å}}^{1600\ \text{Å}} [r_{i,m}(\lambda) - q_i(\lambda)]^2 \, d\lambda}{\int_{1020\ \text{Å}}^{1600\ \text{Å}} [q_i(\lambda) - \mu(\lambda)]^2 \, d\lambda}.$$  

This quantity measures the square of the difference of the reconstructed spectrum from the continuum-fitted QSO spectrum,
| QSO             | z      | $|\delta F_{10}|$ Red Side | $|\delta F_{10}|$ Blue Side | $\delta F_{10}$ Blue Side |
|-----------------|--------|----------------------|-----------------------------|--------------------------|
| Q0003+1553      | 0.450  | 3.3                   | 5.2                         | 4.1                      |
| Q0026+1259      | 0.145  | 5.2                   | 6.7                         | 5.4                      |
| Q0044+0303$^a$  | 0.623  | 4.6                   | 27.3                        | −27.1                    |
| Q0159−1147$^a$  | 0.669  | 2.1                   | 3.1                         | 2.8                      |
| Q0349−1438      | 0.615  | 2.2                   | 14.7                        | −14.5                    |
| Q0405−1219      | 0.572  | 2.3                   | 4.4                         | −4.3                     |
| Q0414−0601      | 0.774  | 3.5                   | 8.7                         | 8.7                      |
| Q0439−4319      | 0.593  | 6.0                   | 5.3                         | 5.3                      |
| Q0454−2203$^a$  | 0.532  | 3.0                   | 31.4                        | −30.6                    |
| Q0624+6907$^a$  | 0.367  | 4.7                   | 29.1                        | −28.8                    |
| Q0637−7513      | 0.652  | 2.7                   | 18.6                        | 18.6                     |
| Q0923+3915      | 0.698  | 2.5                   | 4.5                         | 4.2                      |
| Q0947+3940      | 0.205  | 2.8                   | 15.3                        | 15.5                     |
| Q0953+4129$^a$  | 0.233  | 5.0                   | 16.8                        | 17.4                     |
| Q0954+5537$^a$  | 0.901  | 3.8                   | 2.7                         | 2.1                      |
| Q0959+6827      | 0.767  | 3.5                   | 4.4                         | −2.3                     |
| Q1001+2910$^a$  | 0.328  | 5.3                   | 7.9                         | 7.8                      |
| Q1007+4147      | 0.612  | 3.4                   | 8.8                         | 9.1                      |
| Q1100+7715$^a$  | 0.312  | 3.9                   | 31.0                        | 31.0                     |
| Q1104+1644      | 0.630  | 3.1                   | 2.7                         | −0.4                     |
| Q1115+4042$^a$  | 0.154  | 5.6                   | 14.5                        | 14.7                     |
| Q1137+6604$^a$  | 0.646  | 3.1                   | 2.3                         | −2.3                     |
| Q1148+5454      | 0.970  | 3.0                   | 12.3                        | 12.5                     |
| Q1216+0655      | 0.332  | 9.4                   | 4.7                         | −3.3                     |
| Q1229−2020$^a$  | 1.041  | 3.1                   | 7.0                         | 6.9                      |
| Q1248+4007$^a$  | 1.027  | 4.2                   | 6.4                         | 6.3                      |
| Q1252+1157$^a$  | 0.868  | 3.5                   | 6.8                         | 6.8                      |
| Q1259+5918$^a$  | 0.468  | 3.0                   | 13.5                        | −13.4                    |
| Q1317+2743      | 1.009  | 1.5                   | 10.8                        | 10.8                     |
| Q1320+2925$^a$  | 0.947  | 3.1                   | 4.3                         | 4.2                      |
| Q1322+6557      | 0.168  | 4.8                   | 13.5                        | 13.3                     |
| Q1354+1933      | 0.720  | 4.8                   | 8.1                         | 8.3                      |
| Q1402+2609$^a$  | 0.165  | 3.6                   | 10.6                        | 10.9                     |
| Q1424−1150$^a$  | 0.804  | 4.8                   | 10.6                        | 10.6                     |
| Q1427+4800      | 0.222  | 4.1                   | 25.9                        | 26.1                     |
| Q1444+4047$^a$  | 0.266  | 4.1                   | 15.4                        | −15.1                    |
| Q1538+4745$^a$  | 0.768  | 4.5                   | 23.4                        | −23.2                    |
| Q1544+4855      | 0.400  | 7.4                   | 13.3                        | −12.8                    |
| Q1622+2352$^a$  | 0.926  | 4.5                   | 2.7                         | 0.5                      |
| Q1637+5726      | 0.750  | 2.5                   | 3.6                         | 3.6                      |
| Q1821+6419$^a$  | 0.296  | 5.0                   | 12.9                        | −13.2                    |
| Q1928+7351$^a$  | 0.302  | 2.7                   | 6.1                         | 12.0                     |
| Q2145+0643      | 1.000  | 3.7                   | 2.5                         | 0.4                      |
| Q2201+3131$^a$  | 0.296  | 3.1                   | 3.2                         | 3.3                      |
| Q2243−1222      | 0.626  | 3.6                   | 2.6                         | 2.2                      |
| Q2251+1120$^a$  | 0.325  | 6.8                   | 27.6                        | −27.1                    |
| Q2251+1552      | 0.856  | 2.7                   | 4.1                         | 2.2                      |
| Q2340−0393$^a$  | 0.894  | 3.1                   | 24.7                        | 24.6                     |
| Q2344+0914$^a$  | 0.671  | 7.8                   | 5.5                         | 4.9                      |
| Q2352−3414      | 0.707  | 3.6                   | 3.9                         | 1.0                      |

Note.—We used 10 principal components for the error values, which we show multiplied by 100.

$^a$ The spectrum contains absorption or unusual photon noise in the region 1216–1240 Å that might have led to an inaccurate continuum level.
in units of the square of the difference between that QSO and the mean. Hence, $\delta E_{i,m}$ decreases from 1 to near 0 as we add components to the reconstruction. The $m$ in $\delta E_{i,m}$ tells us that we have used the first $m$ components in the reconstruction $r_{i,\lambda}(\lambda)$. In Table 2 we list the mean $\langle \delta E_{i,m} \rangle = (1/N) \sum_{i=1}^{N} \delta E_{i,m}$, averaged over all 50 QSOs. The first three components account for 77% of the residual, and the first 10 components about 96%. Francis et al. (1992, their Fig. 4) analyzed the LBQS set of QSO spectra (Hewett et al. 1995, 2001). Using a different statistic that accounts for photon noise and different wavelengths and including broad absorption line QSOs and all absorption lines, they found that the first three components accounted for 75% of the variance, and the first 10 components for 95%. We also analyzed the LBQS spectra, kindly provided by Paul Francis and Paul Hewett, to confirm that our implementation of the PCA matched theirs given our wavelength range and selection criteria. The residual variance decreases in a different way for each QSO because the contributions of the components differ. However, on average, the rate of reduction slows after the third component and saturates around the 10th component. The components greater than about the 10th look noisy and carry little information. In the following discussion, we use up to the first 10 principal components.

### 4. PREDICTING SPECTRA

Our goal is to predict the continuum of a QSO in the Ly$\alpha$ forest, the blue side, using the spectrum of wavelengths larger than the Ly$\alpha$ emission, the red side. In § 4.1 we describe how
we relate the blue and red-side continua, and in § 4.2 we give a general recipe to make predictions.

4.1. Methods

Unfortunately, we do not have enough QSO spectra for both a training set and a separate set of spectra that we can use to test our predictions. Hence, we also use the training set for the tests. When we make a prediction for a QSO, we use a set of principal components that we generated without using any spectra of that QSO. We call this the bootstrap method, and we use it in most of the remainder of this paper. When we omit different QSOs, the first five principal components change by a few percent, most noticeably at the wavelengths of the emission lines. If we do not omit the QSO of interest, the complete set of principal components contains all the information on each QSO, including the continuum fitting and photon noise. We found that the noise features could identify a QSO, giving weights with unrealistically high precision.

We chose three steps to quantify the relationship between the red and blue sides of the spectra in the training set:

1. We find the first $m$ principal components, $\xi_j(\lambda)$, and their weights, $c_j$. We use all the blue- and red-side wavelengths, 1020–1600 Å, and we use the bootstrap method to give a slightly different set of $\xi_j(\lambda)$ for each QSO $i$.

2. We repeat step 1, using only the red wavelengths 1216–1600 Å. We again keep the first $m$ principal components, $\xi_j(\lambda)$, and their weights, $d_j$, which are similar to those from step 1. We can write the weights in an $N \times m$ matrix form $C = c_j$, and similarly for $D = d_j$.

3. We find the $m \times m$ projection matrix $X = x_{ij}$, which translates the weights found on the red side to the weights for the whole spectrum:

$$C = D \cdot X.$$
We have \( N = 50 \) QSOs, and we keep \( m = 10 \) components. We then have more equations (\( N \)) than unknowns (\( m \)), and we wish to find the least-squares solution to this overdetermined set of linear equations. The solution matrix \( X \) can be found via the singular value decomposition technique (Press et al. 1992).

### 4.2. Making a Prediction

We are now ready to make a prediction of the Ly\( \alpha \) forest continuum of any QSO spectrum, which need not be in the training set, provided we have the red side of its spectrum. We obtain predictions in three steps, similar to those presented above to find \( X \).

1. We find the weights for the red spectrum,

   \[
   b_{ij} = \int_{1216 \, \text{Å}}^{1600 \, \text{Å}} [g_i(\lambda) - \mu(\lambda)] \zeta_j(\lambda) \, d\lambda. \tag{8}
   \]

   The \( b_{ij} \) are like the \( d_{ij} \) in step 2 in the above list, except that the principal components could be different. If the QSO is not part of the training set, then the \( \zeta_j(\lambda) \) can be derived using the entire training set.

2. We translate the weights from the red side, \( b_{ij} \), to weights for the whole spectrum using

   \[ a_{ij} = \sum_{k=1}^{m} b_{ij} x_{kj}. \tag{9} \]

   This resembles equation (7), except that we now know \( X \) and are deriving the \( a_{ij} \) that play the roles of the \( c_{ij} \) of step 1 in the above list.

3. We make a predicted spectrum,

   \[ p_{i,m}(\lambda) = \mu(\lambda) + \sum_{j=1}^{m} a_{ij} \zeta_j(\lambda). \tag{10} \]

   The predicted spectrum \( p_{i,m}(\lambda) \) differs from the reconstruction \( r_{i,m}(\lambda) \) because the reconstruction uses weights derived from the blue and red sides of the spectrum, using equation (1), whereas the predictions use weights derived from the red part of the spectrum alone, using equations (8) and (9).

   We provide the mean spectrum, the two sets of principal components \( \zeta_j(\lambda) \) and \( \zeta(\lambda) \), and the projection matrix \( X \), so the readers can make their own predictions. Portions of the principal component spectra \( \zeta_j(\lambda) \) and \( \zeta(\lambda) \) are shown in Tables 3 and 4 for guidance. The numerical values of the projection matrix \( X \) are shown in the Appendix. Machine-readable data are published in the electronic edition of the *Astrophysical Journal*.

#### 4.3. Prediction Accuracy

In Figure 3 we show two relatively successful and two unsuccessful predictions, \( p_{i,10}(\lambda) \), together with the corresponding continuum-fitted QSO spectra, \( q_i(\lambda) \). When the prediction fails, the predicted spectrum is systematically either too low or too high, but with no preference.

### TABLE 3

**Principal Component Spectra \( \zeta_j \): 1020 Å \( \leq \lambda \leq \) 1600 Å**

| Wavelength (Å) | Mean \( \mu \) | STD \( \sigma \) | \( \xi_1 \) | \( \xi_2 \) | \( \xi_3 \) | \( \xi_4 \) | \( \xi_5 \) | \( \xi_6 \) | \( \xi_7 \) | \( \xi_8 \) | \( \xi_9 \) | \( \xi_{10} \) |
|----------------|---------------|----------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| 1020.0,......... | 1.4754        | 0.2954         | 0.0036   | -0.0678  | 0.0416   | -0.0132  | 0.0116   | -0.0070  | 0.0547   | -0.0731  | 0.0024   | -0.0213  |
| 1020.5,......... | 1.5006        | 0.3005         | 0.0050   | -0.0681  | 0.0440   | -0.0140  | 0.0121   | -0.0073  | 0.0577   | -0.0739  | -0.0000  | -0.0218  |
| 1021.0,......... | 1.5272        | 0.3060         | 0.0064   | -0.0684  | 0.0465   | -0.0149  | 0.0128   | -0.0076  | 0.0600   | -0.0745  | -0.0024  | -0.0230  |
| 1021.5,......... | 1.5549        | 0.3114         | 0.0078   | -0.0686  | 0.0492   | -0.0158  | 0.0139   | -0.0078  | 0.0620   | -0.0748  | -0.0043  | -0.0245  |
| 1022.0,......... | 1.5838        | 0.3168         | 0.0094   | -0.0686  | 0.0520   | -0.0169  | 0.0152   | -0.0082  | 0.0633   | -0.0753  | -0.0064  | -0.0262  |

**Notes.**—The input 50 spectra were normalized to unity around 1280 Å (\( \xi \)). The complete set of the principal component spectra \( \zeta_j(\lambda) \) is available in its entirety in the electronic edition of the *Astrophysical Journal*. A portion is shown here for guidance regarding its form and content.
We assess the errors on our predictions using the absolute fractional flux error:

\[
|\delta F_{i,m}| = \frac{\int_{\lambda_1}^{\lambda_2} \left| p_i(x) - q_i(x) \right| d\lambda}{\int_{\lambda_1}^{\lambda_2} d\lambda}.
\]

For the blue side, \(\lambda_1 = 1050 \text{ Å} \) and \(\lambda_2 = 1170 \text{ Å} \), avoiding the \(\text{Ly}\beta \) and \(\text{Ly}\alpha \) emission lines, and for the red side, \(\lambda_1 = 1216 \text{ Å} \) and \(\lambda_2 = 1600 \text{ Å} \). In Table 1 we list the absolute fractional flux errors for each QSO when we use \(m = 10\) components. In Table 2, third and fifth rows, we list \(\langle \delta F_{i,m} \rangle = (1/N) \sum_{i=1}^{N} |\delta F_{i,m}|\), the mean of the absolute fractional flux error from all 50 QSOs, and we show how this mean changes with \(m\). For comparison, in the second and fourth rows, we also list \(\langle \delta F_{i,m} \rangle\) obtained by reconstruction when we replace \(q_i(x)\) in equation (11) with \(r_i,m(x)\) from equation (1). These \(r_i,m(x)\) use the principal components derived from the blue plus red sides. We find similar values if we use the principal components from the red side alone.

The advantage of using more components is different for the reconstruction, the red-side prediction, and the blue-side prediction. For the reconstruction and the red-side prediction, more components always improve the fit (Table 2, second through fourth rows), although the reduction of the error is small after about the 10th component, and we think the remaining 3% absolute fractional flux error is similar in size to the error of the continuum fitting (§2).

On the other hand, for the blue-side prediction, adding more components does not reduce the mean error. We think the reason is related to the properties of the first few principal components. Only the third, fourth, and fifth principal components have a significant slope. If we choose appropriate weights for these components, we are likely to make accurate predictions, and adding more components can reduce the residuals (Figs. 3a and 3b). But if we choose inappropriate weights, there is no way to correct the slope (Figs. 3c and 3d). Hence, the main error in our predictions is a systematic slope error that makes the blue-side prediction too high or too low.

Although our predictions give small errors for some QSOs, they give huge errors for others. With 10 components, the mode of the \(\delta F_{i,10}\) distribution is 3%, the median is 8%, and the range is about 30%. We predict the blue continua of 28 out of 50 QSOs to \(|\delta F| \leq 10\%\), but we do not know which QSOs will have these small errors. We find that many of the QSOs that give the largest errors have absorption or unusually low S/N on the red side of the Ly\(\alpha\) line (1216–1240 Å, noted in Table 1), which makes the continuum level difficult to estimate. This region is significant because it contains much of the variance and hence will have a large effect on the weights. However, in others cases there is

| Wavelength (Å) | Mean \(\mu\) | STD \(\sigma\) | \(\zeta_1\) | \(\zeta_2\) | \(\zeta_3\) | \(\zeta_4\) | \(\zeta_5\) | \(\zeta_6\) | \(\zeta_7\) | \(\zeta_8\) | \(\zeta_9\) | \(\zeta_{10}\) |
|---------------|-------------|-------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| 1216.0.........| 5.3986      | 1.9319      | −0.3165   | 0.1868    | −0.0081   | 0.3034    | −0.3835   | 0.0815    | −0.0672   | 0.1249    | −0.1830   | 0.0337    |
| 1216.5.........| 5.2560      | 1.8206      | −0.3040   | 0.1636    | −0.0125   | 0.2717    | −0.2622   | 0.0673    | −0.0355   | 0.0662    | −0.0308   | −0.0214   |
| 1217.0.........| 5.0261      | 1.6645      | −0.2831   | 0.1273    | −0.0174   | 0.2202    | −0.1053   | 0.0327    | 0.0273    | −0.0013   | 0.1267    | −0.0563   |
| 1217.5.........| 4.7976      | 1.5134      | −0.2599   | 0.0930    | −0.0208   | 0.1757    | −0.0011   | −0.0045   | 0.1019    | −0.0360   | 0.1885    | −0.0536   |
| 1218.0.........| 4.5331      | 1.3678      | −0.2366   | 0.0665    | −0.0250   | 0.1337    | 0.0721    | −0.0243   | 0.1435    | −0.0361   | 0.1667    | −0.0182   |

Notes.—As for Table 3, the input 50 spectra were normalized to unity around 1280 Å (§2). The complete set of the principal component spectra \(\zeta(x)\) is available in its entirety in the electronic edition of the Astrophysical Journal. A portion is shown here for guidance regarding its form and content.
little or no absorption, and the slope of the spectrum appears to change as we cross the \( \text{Ly}\alpha \) emission line. We do not in general see any correlation between the errors in the red- and blue-side predictions.

We introduce a third error statistic, the fractional flux error,

\[
\delta F_{i,m} = \frac{\int_{\lambda_i}^{\lambda_{i+1}} \left\{ \left[ p_i(\lambda) - g_i(\lambda) \right] / q_i(\lambda) \right\} \, d\lambda}{\int_{\lambda_i}^{\lambda_{i+1}} d\lambda},
\]

(12)

to help us estimate the error in the integrated flux. This error is related to the flux decrement statistic \( D_{\lambda} \) (Oke & Korycansky 1982; Bernardi et al. 2003), which has been used to calibrate numerical simulations of the IGM (Croft et al. 2002). In Table 1 we see that \( \delta F_{i,10} \) has a magnitude similar to \( \langle \delta F_{i,10} \rangle \), also for the blue side, because the predictions tend to be too low or too high across the entire blue side. At each wavelength, the mean of \( \delta F_{i,m} \) for all 50 QSOs, \( \langle \delta F_{i,m} \rangle \), is 0 by definition when we use \( m = 0 \), because \( p_i(\lambda) = \mu(\lambda) \). When \( m > 0 \), the statistic \( \langle \delta F_{i,m} \rangle \) represents the bias in the predictions, and in Table 2 we see that it remains within a few percent of 0 for all \( m \leq 10 \).

In Figure 4 we show an example of the prediction of the \( \text{Ly}\alpha \) forest continuum of a higher redshift QSO from the Sloan Digital Sky Survey (York et al. 2000; Stoughton et al. 2002), for which the continuum is unobservable because of the large amount of absorption. To make this prediction, which looks acceptable, we use the red side of its spectrum.

4.4. Errors and Different Methods

Although the error in the levels of many of the predictions was unexpected, we can think of many possible explanations, including the intrinsic QSO spectra, calibration errors, and the method. The intrinsic slope of the continuum may be changing at the UV wavelengths we consider (Zheng et al. 1997; Telfer et al. 2002). The UV flux from QSOs is blackbody radiation from the accretion disk around the central supermassive black hole (Malkan 1983; Sun & Malkan 1989). If we pass over the peak of the blackbody continuum, the slope changes rapidly, and we may not have enough wavelength coverage to follow the change of the slope.

There are many possible errors in the calibration of the spectra, especially the Galactic extinction correction (Fitzpatrick 1999). At our wavelengths, the extinction increases rapidly as the wavelength drops. If we underestimate color excess \( E(B-V) \) for a \( z = 0 \) QSO by 0.01 mag, we decrease the flux at 1600 \( \AA \) by 7%, and at 1020 \( \AA \) by 14%.

We have presented only one of many ways of predicting spectra. We attempted some other less successful schemes before we arrived at this method. For instance, we fitted the red side by minimizing the \( \chi^2 \) of \( g_i(\lambda) \) against \( a_{ij} \), but we found that the blue-side prediction was then unstable. Different combinations of the principal components give similar \( \chi^2 \) on the red side, but their blue sides can have a large variety.

We also experimented with different ways of normalizing the spectra, both near 1450 \( \AA \) and using the entire red side, and we obtained similar results. We tried attempting to remove the slopes from the spectra by fitting a straight line to the red sides, but these lines were not adequately determined.

The predictions given here could be improved in several ways. We should first seek improved methods, which might not involve PCA. If we had spectra of many more QSOs we would be less subject to the distortions and noise in individual spectra, and we could test the predictions on spectra that were not in the training set. Higher resolution and higher S/N spectra would help to reveal the weak emission features and reduce the errors in the initial continuum fit. Extending the wavelength range on the red side may help identify the slope of the continuum for some QSOs, but for others we find that a reduced wavelength range, from 1216 to 1400 \( \AA \), gives better predictions, because this restricted range has a stronger correlation with the flux on the blue side. We have also made predictions with the red side restricted to 1280–1600 \( \AA \), to avoid the large flux errors that can occur when there are absorption lines in the \( \text{Ly}\alpha \) and \( N\,\lambda \) emission lines. We find that there was no significant change in \( \langle |\delta F_{i,m}| \rangle \).

This work was funded primarily by NASA grant NAG5-9224, in part by NAG5-13113, and also by NSF grants AST 99-00842 and AST 00-98731. We are especially grateful to Paul Francis, who provided the code and LBQS spectra that he had analyzed in his 1992 paper and who answered many questions. Paul Hewett kindly sent the error arrays for those spectra. Wei Zheng and Buell Januzzi kindly provided copies of \( HST \) QSO spectra that we used before we located the invaluable collection of \( HST \) spectra posted to the web by Jill Bechtold. N. S. thanks the Okamura Group at the University of Tokyo from whom he learned the basics of PCA. D. T. is grateful for the hospitality of the National Astronomical Observatory of Japan in Mitaka, where this work was begun during a visit in 1999. He was sponsored by the Ministry of Education, Science, Sports, and Culture. We especially thank Masanori Iye for his hospitality and advice with the early part of this project. We thank Carl Melis and Angela Chapman for their careful reading of this manuscript. We thank the referee, Gerard Williger, for constructive suggestions.
The following is the $10 \times 10$ projection matrix $X$ discussed in §4:

$$X = \begin{pmatrix}
-1.326 & -0.042 & -0.000 & -0.040 & -0.016 & -0.011 & -0.001 & -0.004 & 0.000 & 0.001 \\
0.023 & 1.158 & 0.054 & -0.488 & 0.078 & -0.120 & -0.016 & 0.013 & 0.003 & -0.006 \\
-0.143 & 0.457 & 0.883 & 0.528 & -0.251 & -0.224 & -0.115 & 0.001 & 0.014 & -0.016 \\
0.484 & -0.405 & 0.348 & -0.705 & -0.829 & 0.255 & 0.186 & -0.015 & -0.008 & -0.034 \\
-0.112 & 0.053 & -0.241 & 0.335 & -0.013 & -0.680 & 0.973 & 0.249 & 0.207 & -0.197 \\
-0.596 & 1.105 & 0.430 & 0.419 & 0.224 & 1.162 & 0.310 & 0.025 & 0.517 & 0.185 \\
-0.628 & 1.206 & -0.806 & 0.704 & -0.468 & 0.563 & -0.181 & 0.679 & -0.292 & -0.317 \\
0.414 & -0.301 & -1.021 & -0.112 & -0.552 & -0.529 & -0.665 & -0.141 & 0.840 & -0.249 \\
1.667 & 0.372 & -0.254 & 0.213 & -0.082 & -0.506 & 0.264 & 0.327 & -0.490 & 0.445 \\
3.039 & 0.344 & -0.426 & -0.321 & -0.515 & -0.585 & -0.139 & 0.118 & -0.080 & -0.144
\end{pmatrix}$$

REFERENCES

Bahcall, J. N., et al. 1993, ApJS, 87, 1
———. 1996, ApJ, 457, 19

Bechtold, J., Dobrzycki, A., Wilden, B., Morita, M., Scott, J., Dobrzycka, D., Tran, K.-V., & Aldcroft, T. L. 2002, ApJS, 140, 143

Becker, R. H., et al. 2001, AJ, 122, 2850

Bernardi, M., et al. 2003, AJ, 125, 32

Burstein, D., & Heiles, C. 1982, AJ, 87, 1165

Cardelli, J. A., Clayton, G. C., & Mathis, J. S. 1989, ApJ, 345, 245

Croft, R. A. C., Weinberg, D. H., Bolte, M., Burles, S., Hernquist, L., Katz, N., Kirkman, D., & Tytler, D. 2002, ApJ, 581, 20

Djorgovski, S. G., Castro, S., Stern, D., & Mahabal, A. A. 2001, ApJ, 560, L5

Efstathiou, G. 2002, MNRAS, 332, 193

Fan, X., et al. 2001, AJ, 122, 2833

Fitzpatrick, E. L. 1999, PASP, 111, 63

Francis, P. J., Hewett, P. C., Foltz, C. B., & Chaffee, F. H. 1992, ApJ, 398, 476

Hewett, P. C., Foltz, C. B., & Chaffee, F. H. 1995, AJ, 109, 1498
———. 2001, AJ, 122, 518

Jannuzi, B. T., et al. 1998, ApJS, 118, 1

Kanbur, S. M., Iono, D., Tanvir, N. R., & Hendry, M. A. 2002, MNRAS, 329, 126

Kirkman, D., Tytler, D., Suzuki, N., O’Meara, J. M., & Lubin, D. 2003, ApJS, 149, 1

Malkan, M. A. 1983, ApJ, 268, 582

Oke, J. B., & Korycansky, D. G. 1982, ApJ, 255, 11

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 1992, Numerical Recipes in FORTRAN (2nd ed.; Cambridge: Cambridge Univ. Press)

Stoughton, C., et al. 2002, AJ, 123, 485

Sun, W., & Malkan, M. A. 1989, ApJ, 346, 68

Suzuki, N., Tytler, D., Kirkman, D., O’Meara, J. M., & Lubin, D. 2003, PASP, 115, 1050

Telfer, R. C., Zheng, W., Kriss, G. A., & Davidsen, A. F. 2002, ApJ, 565, 773

Vanden Berk, D. E., et al. 2001, AJ, 122, 549

White, R. L., Becker, R. H., Fan, X., & Strauss, M. A. 2003, AJ, 126, 1

Whitney, C. A. 1983, A&AS, 51, 443

Wills, B. J., et al. 1995, ApJ, 447, 139

York, D. G., et al. 2000, AJ, 120, 1579

Zheng, W., Kriss, G. A., Telfer, R. C., Grimes, J. P., & Davidsen, A. F. 1997, ApJ, 475, 469