The Astronomical Journal, 143:51 (14pp), 2012 February doi:10.1088/0004-6256/143/2/51

MEAN-FLUX-REGULATED PRINCIPAL COMPONENT ANALYSIS CONTINUUM FITTING OF SLOAN DIGITAL SKY SURVEY Lyα FOREST SPECTRA

Khee-Gan Lee[1], Nao Suzuki[2], and David N. Spergel[1]
[1] Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544, USA; lee@astro.princeton.edu
[2] E.O. Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA
Received 2011 August 29; accepted 2011 December 7; published 2012 January 13

ABSTRACT

Continuum fitting is an important aspect of Lyα forest science, since errors in the derived optical depths scale with the fractional continuum error. However, traditional methods of estimating continua in noisy and moderate-resolution spectra (e.g., Sloan Digital Sky Survey, SDSS; S/N $\lesssim$ 10 pixel^{-1} and $R \sim 2000$), such as power-law extrapolation or dividing by the mean spectrum, achieve no better than ∼15% rms accuracy. To improve on this, we introduce mean-flux-regulated principal component analysis (MF-PCA) continuum fitting. In this technique, PCA fitting is carried out redward of the quasar Lyα line in order to provide a prediction for the shape of the Lyα forest continuum. The slope and amplitude of this continuum prediction is then corrected using external constraints for the Lyα forest mean flux. This requires prior knowledge of the mean flux, $\langle F \rangle$, but significantly improves the accuracy of the flux transmission, $F \equiv \exp(-\tau)$, estimated from each pixel. From tests on mock spectra, we find that MF-PCA reduces the errors to 8% rms in S/N $\sim$ 2 spectra, and <5% rms in spectra with S/N $\gtrsim$ 5. The residual Fourier power in the continuum is decreased by a factor of a few in comparison with dividing by the mean continuum, enabling Lyα flux power spectrum measurements to be extended to ∼2× larger scales. Using this new technique, we make available continuum fits for 12,069 $z > 2.3$ Lyα forest spectra from SDSS Data Release 7 for use by the community. This technique is also applicable to future releases of the ongoing Baryon Oscillations Spectroscopic Survey, which obtains spectra for ∼150,000 Lyα forest spectra at low signal-to-noise (S/N ∼ 2).

Key words: intergalactic medium – methods: data analysis – quasars: absorption lines – quasars: emission lines

Online-only material: color figures

1. INTRODUCTION

Over the past two decades, the Lyα forest absorption observed in the spectrum of high-redshift quasars has been an important probe of large-scale structure and the intergalactic medium (IGM) at $z \gtrsim 2$. The fundamental quantity of interest of the Lyα forest is its local optical depth to absorption, $\tau(\vec{x})$, at position $\vec{x}$. This is not a directly observed quantity: It is derived from the flux transmission $F \equiv e^{-\tau}$, which requires knowledge of the intrinsic quasar continuum $C(\lambda_{\rm rest})$ in order to be extracted from the observed flux, $F \equiv f(\lambda_{\rm obs})/C(\lambda_{\rm rest})$, where $f(\lambda_{\rm obs})$ is the observed flux and $\lambda_{\rm rest} \equiv \lambda_{\rm obs}/(1 + z_{\rm QSO})$ is the rest-frame wavelength of the quasar at redshift $z = z_{\rm QSO}$. The error in the measured optical depth, $\delta\tau$, roughly scales with the fractional continuum error, $\delta\tau \sim \delta C/C$, which means that accurate continuum fitting is important in the optically thin Lyα forest, where $\tau \lesssim 1$. Therefore, accurate estimates of the underlying quasar continuum are crucial in order to take full advantage of modern Lyα forest data sets.

Virtually, all aspects of Lyα forest science are dependent on the continuum determination. Studies of the IGM ionizing background, for example, require accurate measurements of the mean flux, $\langle F \rangle = \exp(-\tau)$, of the Lyα forest (see, e.g., Tytler et al. 2004; Bolton et al. 2005; Kirkman et al. 2005; Faucher-Giguère et al. 2008, 2009). Measurements of $\langle F \rangle$ are sensitive to the *bias* of the continuum estimation, but not so much on the *random error* of each measured pixel since, by definition, many Lyα forest pixels need to be averaged in order to make the measurement at each redshift bin. Higher-order statistics are more susceptible to continuum errors propagating into the flux transmissions estimated from each individual data pixel, such as the flux probability distribution function (Desjacques et al. 2007; Lee & Spergel 2011), flux power spectrum (Croft et al. 1999, 2002; McDonald et al. 2000, 2006), and bispectrum (Mandelbaum et al. 2003; Viel et al. 2004). For example, the unaccounted continuum variance in the quasar continuum has limited the measurement of the one-dimensional Lyα forest flux power spectrum (McDonald et al. 2006) to comoving scales of $r \lesssim 40\,h^{-1}$ Mpc at $z = 2.5$. Ongoing attempts to measure the baryon acoustic oscillation (BAO) feature from the Lyα forest using transverse correlations across lines of sight are less sensitive to continuum errors (McDonald & Eisenstein 2007); however, large continuum errors will still degrade the significance of the BAO signal measured in this fashion.

Unfortunately, accurate quasar continuum fitting is a non-trivial problem. At redshifts ($z \gtrsim 2$) in which the Lyα forest becomes accessible to ground-based optical telescopes, the high absorber line density makes it challenging to identify the intrinsic quasar continuum. With high-resolution and high signal-to-noise (S/N) quasar spectra obtained from large telescopes, continua are usually fitted using some form of spline fitting to the observed transmission peaks of the forest—if one believes that the transmission peaks truly reach the quasar continuum at a given redshift (but see Faucher-Giguère et al. 2008; Lee 2011). These direct fitting methods cannot, in general, be applied to large data sets such as the Sloan Digital Sky Survey (SDSS), as the modest resolution and low-S/N make it impossible to directly fit the Lyα forest except for the very highest S/N subsamples—and even then steps need to be taken to account for the degradation of transmission peaks from the lower resolution (see, e.g., Dall'Aglio et al. 2009). Moreover, direct fitting techniques usually require significant human intervention and are often very time consuming, precluding their application to the $\sim 10^4$ Lyα forest sightlines in the SDSS.

Noisy Lyα forest data usually require some form of extrapolation from the relatively unabsorbed spectrum blueward[3] of the quasar's Lyα λ1216 broad emission line. The simplest way to do this is to fit a power law $f_\nu \propto \nu^\alpha$ to spectral regions uncontaminated by quasar emission lines redward of Lyα, or even to eschew fitting individual spectra and extrapolate a mean power law from $\lambda_{\rm rest} > 1216$ Å, using power-law values from the literature (e.g., from Vanden Berk et al. 2001).

There are two major issues with power-law estimation of Lyα forest continua. First, there is a break in the underlying quasar power law at $\lambda_{\rm rest} \sim 1200$–$1300$ Å. This was first identified by Zheng et al. (1997) from a study of low-redshift ($z_{\rm QSO} \sim 1$) quasars observed in the ultraviolet in which the Lyα forest continuum could be clearly identified due to the low Lyα line density at those epochs. A subsequent study by Telfer et al. (2002) found mean power-law indices of $\langle\alpha_{\rm NUV}\rangle = -0.69$ and $\langle\alpha_{\rm EUV}\rangle = -1.76$ redward and blueward of $\lambda_{\rm rest} \sim 1200$ Å, respectively. This implies that a naïve power-law extrapolation from $\lambda_{\rm rest} > 1216$ Å would underestimate the true Lyα forest continuum by ~10%. Furthermore, Telfer et al. (2002) found a large scatter in $\alpha_{\rm NUV}$ and $\alpha_{\rm EUV}$ from the individual quasars in their sample, with no correlation between the two; this increases the error from power-law extrapolation in individual Lyα forest spectra. While Desjacques et al. (2007) and Pâris et al. (2011) have discussed this EUV–NUV power-law break in the context of Lyα forest continuum estimation, it is often ignored in Lyα forest analyses.

Second, the Lyα forest "continuum" (usually defined around $\lambda_{\rm rest} \approx 1040$–$1180$ Å) includes weak emission lines such as Fe II λ1071 and Fe II/Fe III λ1123, although the exact identifications vary from author to author. These emission lines can cause deviations of up to ~10% from a flat continuum. It is possible to take these features into account on average: For example, Bernardi et al. (2003) modeled them as two Gaussian functions superposed on top of an underlying power law. However, there is a great diversity in the shape and equivalent width of these weak emission lines (Suzuki 2006). Therefore, the use of an average continuum shape would not account for variations of up to 10% within individual quasars due the presence of these emission lines.

One possible avenue for improved quasar continuum fits is principal component analysis (PCA). Suzuki et al. (2005) explored this using a sample of 50 low-redshift quasars observed in the UV by the *Hubble Space Telescope* (*HST*) in which the $\lambda_{\rm rest} < 1216$ Å continuum can be clearly identified. They concluded that while PCA fits to the red side ($\lambda_{\rm rest} = 1216$–$1600$ Å) of an individual spectrum gave a good prediction of the Lyα continuum shape (i.e., the weak emission lines), the overall continuum amplitude had ~10% errors. This is presumably due to the EUV–NUV power law break. Pâris et al. (2011) recently carried out a similar analysis on a high-S/N (S/N $\gtrsim 10$ pixel^{-1}) subsample of the SDSS quasar sample. They found a better prediction accuracy of ~5%, possibly due to their larger spectral baseline ($\lambda_{\rm rest} \approx 1025$–$2000$ Å as opposed to $\lambda_{\rm rest} \approx 1025$–$1600$ Å in the earlier work).

However, the standard PCA formalism does not take pixel noise into account, whereas the majority of quasars observed in the SDSS have low signal-to-noise (S/N < 10). Francis et al. (1992) have argued that PCA fitting errors scale directly with the noise level, which implies that, e.g., one can expect no better than ~20% continuum accuracy in an S/N = 5 spectrum using PCA, even without taking the power-law break into account.

For all the reasons outlined above, more accurate continuum-fitting methods are sorely needed to take full advantage of Lyα forest data from SDSS and future spectroscopic surveys. In this paper, we will explore a refinement of the PCA technique that we term "mean-flux-regulated PCA" (MF-PCA). Briefly, we carry out least-squares fitting of PCA templates to the unabsorbed quasar spectrum redward ($\lambda_{\rm rest} > 1216$ Å) of the Lyα line in order to obtain a prediction for the continuum shape, and then use the expected mean flux, $\langle F\rangle(z)$, to constrain the amplitude of the fitted continuum. Tytler et al. (2004) have shown that the dispersion in $\langle F\rangle$ expected from a $\Delta z = 0.1$ segment of the Lyα forest at $z = 2$ is $\sigma_F(\Delta z = 0.1) \approx 4\%$. When averaged across an entire Lyα forest sightline (which spans $\Delta z \approx 0.4$ for a quasar at $z_{\rm QSO} = 2.5$), one expects the continuum amplitude to be predicted to ~2%.

Note that this method requires some prior assumptions about the mean flux of the Lyα forest, $\langle F\rangle(z)$, which we can obtain from the existing literature. This clearly means that our method cannot be used to measure $\langle F\rangle(z)$, which is unfortunate since it is an important aspect of Lyα forest studies. However, through this compromise we will dramatically reduce the level of continuum errors in individual sightlines, even in very noisy (S/N ~ 2) spectra. This will significantly improve the utility of noisy Lyα forest data for higher-order Lyα forest statistics, such as the flux PDF, flux power spectrum, correlation functions, etc.

This paper is organized as follows. Section 2 describes the publicly available SDSS quasar sample which will be the initial subject of our new technique. Section 3 elucidates the MF-PCA technique, which is then tested on mock spectra in Section 4. We will then discuss the results of our continuum fitting and future improvements. The continuum fits have been made publicly available and can be downloaded via anonymous FTP.[4]

2. DATA

The MF-PCA technique that we develop in this paper is optimized toward large sets of noisy Lyα forest data spectra. We will apply this technique to the SDSS data, which comprises ~10^4 Lyα forest sightlines at moderate resolution ($R \approx 2000$) and modest signal-to-noise ratio (S/N ~ few per pixel).

This section provides an overview of the SDSS Lyα forest data sample and also the two sets of quasar templates that we will use to fit this data set.

2.1. SDSS DR7 Lyα Forest Sample

In this paper, we carry out continuum fitting for publicly available spectra from the final SDSS Data Release 7 (DR7) quasar catalog (Schneider et al. 2010), which is comprised of 105,783 spectroscopically confirmed quasars observed from the 2.5 m SDSS Telescope in Apache Point, NM. The spectra cover the observed wavelength range $\lambda_{\rm obs} = 3800$–9200 Å with a spectral resolution of $R \equiv \lambda/\Delta\lambda \approx 2000$.

From this overall catalog, we select a subsample suitable for Lyα forest studies. First, we require that some portion of the quasar Lyα forest region, $\lambda_{\rm rest} = 1041$–$1185$ Å, be within the observed wavelength range. Since the extreme blue end (near $\lambda_{\rm obs} \approx 3800$ Å) of the SDSS spectra are known to suffer from spectrophotometric problems, we use $\lambda_{\rm obs} = 3840$ Å as the lower wavelength limit. This sets a minimum quasar redshift of

[3] In this paper, we use the terms "blue" and "red" relative to the quasar Lyα emission line unless otherwise noted.

[4] ftp.astro.princeton.edu/lee/continua/

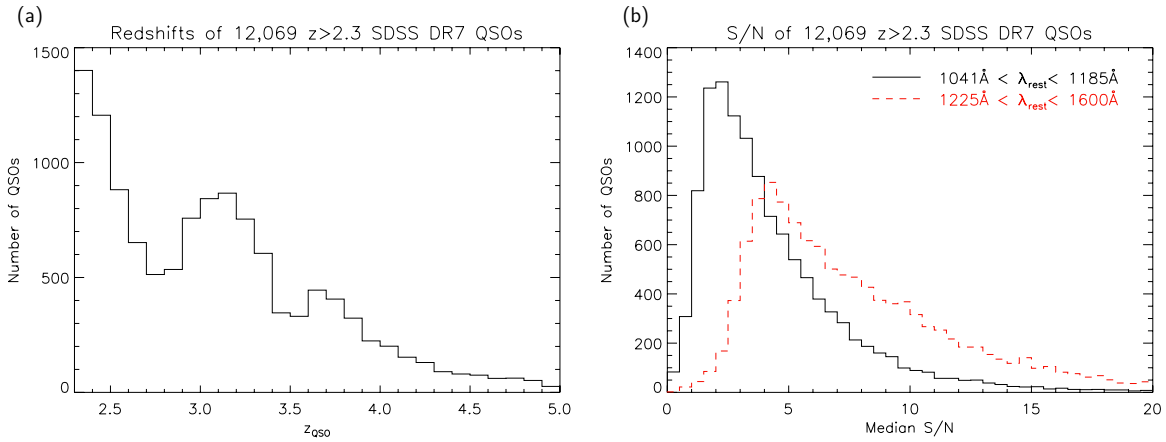

Figure 1. (a) Redshift distribution of the SDSS DR7 quasars fitted in this paper, in $\Delta z_{QSO} = 0.1$ bins. We have selected objects with $z_{QSO} \geqslant 2.3$, which have reasonable coverage of the Lyα forest. Fifty-three quasars with $z_{QSO} > 5.0$ are not shown in this plot. The gaps at $z_{QSO} \approx 2.7$ and $z_{QSO} \approx 3.5$ are where the quasar colors cross the stellar locus, making it difficult to select quasars with these redshifts (see, e.g., Richards et al. 2002). (b) Median S/N per pixel in our quasar sample, in the Lyα forest ($\lambda_{rest} = 1041$–1185 Å; black solid lines) and redward of the quasar Lyα emission line ($\lambda_{rest} = 1225$–1600 Å; red dashed lines). The histograms are in bins of $\Delta S/N = 0.5$. Note that the majority of the Lyα forest sightlines have S/N < 10 pixel^{-1}.

(A color version of this figure is available in the online journal.)

$z_{QSO} = 2.3$. For the quasars that satisfy this redshift criterion, we excise the portions of the spectra that lie below $\lambda_{obs} = 3840$ Å. In addition, broad absorption line (BAL) quasars have continua that are difficult to characterize (although see Allen et al. 2011 for a method to recover quasar continua from BALs), therefore we discard quasars flagged as BALs in the Shen et al. (2011) value-added quasar catalog.

There are 13,133 quasars in the DR7 quasar catalog that satisfy the above criteria. We make further quality cuts by discarding 962 spectra that have SPPIXMASK = 0–12 bitmask set (this signifies issues with the fiber; see Stoughton et al. 2002 for further details on the SDSS bitmask system) and two spectra where the S/N was too low to normalize the spectra, leading to negative normalizations. This leaves us with a sample of 12,069 spectra to which we will apply the MF-PCA technique. Within individual spectra, we mask pixels which have either zero inverse variance or the SPPIXMASK = 16–28 bitmask set. This avoids the use of problematic pixels, such as rejected extractions, bright sky-lines, or bad flats.

The median S/N in the sample is S/N = 3.0 per 69 km s^{-1} SDSS pixel within the Lyα forest, and S/N = 6.2 pixel^{-1} in the $\lambda_{rest} = 1225$–1600 Å wavelength region. The redshift and S/N[5] distributions of our final quasar sample is shown in Figure 1. It is clear that the Lyα forest data from DR7 are noisy. Most of the sightlines have median S/N < 10 within the Lyα forest, which is too noisy to be fitted individually using existing techniques.

In addition, we need to deal with Damped Lyα Absorbers (DLAs) within the spectra. These are absorbing systems with neutral hydrogen column densities of $N_{H\,I} \geqslant 2 \times 10^{20}$ cm^{-2} that result in complete absorption over large portions ($\Delta v \sim 10^3$ km s^{-1}) of affected sight lines. Since the MF-PCA technique (Section 3) fits the amplitude of the quasar continuum based on the mean flux of the low column-density Lyα forest, the excess absorption of a DLA within a sightline would bias the continuum estimate.

To correct for this, we use a catalog of 1427 DLAs identified in the SDSS DR7 spectra by Noterdaeme et al. (2009). First, we mask the wavelength region corresponding to the equivalent width of each DLA (Draine 2011):

$$W \approx \lambda_\alpha \left[\frac{e^2}{m_e c^2} N_{H\,I} f_\alpha \lambda_\alpha \left(\frac{\gamma_\alpha \lambda_\alpha}{c} \right) \right]^{1/2}, \quad (1)$$

where $\lambda_\alpha = 1216$ Å is the rest-frame wavelength of the hydrogen Lyα transition, e is the electron charge, m_e is the electron mass, c is the speed of light, $N_{H\,I}$ is the H I column density of the DLA, f_α is the Lyα oscillator strength, and γ_α is the sum of the Einstein A coefficients for the transition.

However, the damping wings of each DLA extend beyond the equivalent width, providing a small but non-negligible excess absorption to the pixels close to the DLA. We correct for this by multiplying each pixel in the spectrum with $\exp(\tau_{wing}(\Delta\lambda))$, where

$$\tau_{wing}(\Delta\lambda) = \frac{e^2}{m_e c^2} \frac{\gamma_\alpha \lambda_\alpha}{4\pi} f_\alpha N_{H\,I} \lambda_\alpha \left(\frac{\lambda}{\Delta\lambda} \right)^2 \quad (2)$$

and $\Delta\lambda \equiv \lambda - \lambda_\alpha$ is the wavelength separation in the DLA rest frame.

2.2. Quasar Templates

In order to predict the shape of the Lyα forest continuum, we need a set of template spectra with clearly identified continua at wavelengths $\lambda_{rest} < 1216$ Å. For this purpose, we will use two different sets of quasar templates, derived from quasars observed in the *HST*, and SDSS itself.

Suzuki et al. (2005) derived PCA templates from 50 quasars that had been observed by the Far Object Spectrograph (FOS) on the *HST* in the ultraviolet. At the low redshifts ($0.14 < z_{QSO} < 1.04$) of these quasars, the line density of the Lyα forest is sufficiently small that the quasar continuum could be clearly identified. This enabled the creation of templates in the range $\lambda_{rest} = 1025$–1600 Å, covering Lyβ λ1025 to C IV λ1549.

Pâris et al. (2011) recently carried out a similar study, applying the techniques in Suzuki et al. (2005) to a subsample of 78 SDSS DR7 quasars. These $z_{QSO} \approx 3$ quasars were selected to have full coverage of the Lyα forest and relatively high S/N ($\gtrsim 10$ pixel^{-1}). The transmission peaks in the Lyα

[5] Henceforth, all signal-to-noise values quoted in this paper are the median values per 69 km s^{-1} pixel, in the range $\lambda_{rest} = 1225$–1600 Å unless indicated otherwise.

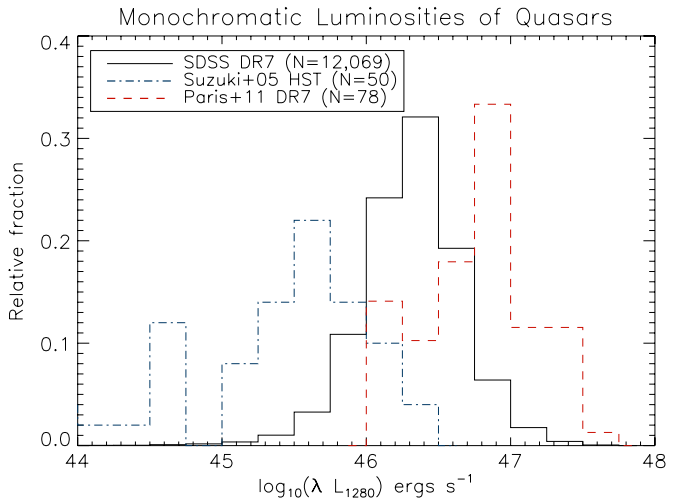

Figure 2. Intrinsic luminosity distribution of quasars from SDSS DR7 (12,069 spectra; black solid line), Suzuki et al. (2005; 50 spectra; blue dot-dashed line), and Pâris et al. (2011; 78 spectra; red dashed line), as estimated from λL_{1280}. These histograms have bin widths of $\Delta \log_{10}(\lambda L_{1280}) = 0.25$ and are normalized such that the sum of all the bins in each histogram is unity.

(A color version of this figure is available in the online journal.)

forest were hand fitted with a low-order spline function to provide a continuum estimate. PCA templates were then derived in the spectral range $\lambda_{\rm rest} = 1020$–$2000$ Å, which included the C III λ1906 line. While this process might give a biased continuum level due to the low resolution and S/N of the templates, it should provide a good description of the relative shape of the quasar continuum which is required for MF-PCA—the mean-flux regulation process is designed to correct for uncertainties in the overall continuum level arising from pure PCA fitting.

We do not expect the redshift differences between the template and the DR7 quasars to be a significant issue, even for the Suzuki et al. (2005) quasars ($\langle z_{\rm QSO} \rangle \approx 0.6$). This is because various studies (e.g., vanden Berk et al. 2004; Fan 2006) have suggested that there is little redshift evolution of quasar spectra. However, the shape of quasar spectra is known to have a significant luminosity dependence, such as the well-known Baldwin effect (Baldwin et al. 1978), which is the anti-correlation between the strength of the C IV λ1549 emission line and the quasar luminosity. It is therefore reasonable to suppose that there would be a significant difference in the spectral shapes represented by two templates and the overall SDSS sample. This is because the Suzuki et al. (2005) sample is comprised of relatively low-luminosity, nearby quasars as opposed to the Pâris et al. (2011) quasars, which were selected to have high S/N and are therefore a luminous subsample of the SDSS quasars.

In order to compare the relative luminosities, we calculate λL_{1280}, the intrinsic monochromatic luminosity near $\lambda_{\rm rest} = 1280$ Å, for the quasars in our SDSS DR7 sample as well as the two template samples. We assume a standard ΛCDM cosmology with $h = 0.7$, $\Omega_m = 0.28$, and $\Omega_m + \Omega_\Lambda = 1$. The respective distributions of λL_{1280} are shown in Figure 2. The SDSS DR7 quasars have a typical luminosity of $\lambda L_{1280} \approx 10^{46.3}$ erg s^{-1}, while the Suzuki et al. (2005) and Pâris et al. (2011) quasars are about 0.5 dex fainter and brighter, respectively. However, the combined luminosity distributions of the Suzuki et al. (2005) and Pâris et al. (2011) template quasars significantly overlap the full range of SDSS DR7 quasars, which justifies the use of both templates in this paper.

3. METHOD

The MF-PCA fitting method described in this paper is essentially a two step process. The first step is (1) fitting of the red side ($\lambda_{\rm rest} > 1216$ Å) of the individual quasar spectra using PCA templates to predict the shape of the weak emission lines in the Lyα forest continuum. This is followed by (2) constraining the amplitude of the predicted Lyα forest continuum to be consistent with existing measurements of the mean-flux evolution of the Lyα forest, $\langle F \rangle(z)$.

3.1. Least-squares PCA Fitting

The basic concept of PCA is that a normalized quasar spectrum, $f(\lambda)$, can be represented as

$$f(\lambda) \approx \mu(\lambda) + \sum_{j=1}^{m} c_j \xi_j(\lambda), \quad (3)$$

where $\mu(\lambda)$ is the mean quasar spectrum, $\xi_j(\lambda)$ is the jth principal component or "eigenspectrum," and c_j are the weights for an individual quasar. The formalism for deriving the eigenspectra and weights is described in Suzuki et al. (2005) and Pâris et al. (2011).

The standard PCA formalism for deriving the weights, c_j, does not take into account spectral noise, which renders it unsuitable for noisy SDSS spectra (see Figure 1(b)). Instead, we first carry out a least-squares fit to the red side of each spectrum using the full $\lambda_{\rm rest} \approx 1000$–$1600$ Å eigenspectra as a basis. Due to the correlation between the weak emission lines within the Lyα forest and in $\lambda_{\rm rest} \sim 1300$–$1500$ Å (Suzuki et al. 2005), we expect this to provide a reasonable prediction for the shape of the continuum.

As described in Section 2.2, we have two separate sets of PCA eigenspectra from Suzuki et al. (2005) and Pâris et al. (2011). In principle, one could combine the two sets of quasar templates to generate one set of eigenspectra that would encompass the diversity of both template samples. However, the template spectra from Pâris et al. (2011) are not available to us at the time of writing, therefore we will carry out our fitting procedure separately for the two sets of PCA eigenspectra. Suzuki et al. (2005) had found that out of their 10 principal component eigenspectra, only the first eight components appeared to describe physical features in the spectra, while the ninth and tenth components seemed to describe mostly noise. Therefore, we will use only eight components from each set of eigenspectra for our fits. In addition, for the sake of consistency we limit ourselves to the rest-wavelength range $\lambda_{\rm rest} = 1020$–$1600$ Å of each eigenspectrum even though the Pâris et al. (2011) eigenspectra extend up to $\lambda_{\rm rest} = 2000$ Å.

However, we have found that fitting the SDSS spectra with just the PCA weights c_j was insufficient to account for the large diversity of the sample. Therefore, we introduce two additional fit parameters: a power-law component, α_λ, and redshift-correction factor, c_z. The power-law component, α_λ, is necessary due to the large range of slopes found in the SDSS quasars. Even though the third through fifth principal components in the *HST* eigenspectra include the spectral slope, they also describe some emission-line features—the introduction of α_λ as a free parameter allows an additional degree of freedom and enables a better fit to the emission lines and slope simultaneously. Because of this degeneracy between the slopes within the eigenspectra and α_λ, we do not interpret the latter as the slope of the underlying

Table 1
Free Parameters in MF-PCA Continuum Fits

Fit Parameter	Description
f_{1280}	Flux normalization, evaluated at $\lambda_{\rm rest} \approx 1280$ Å
c_z	Redshift-correction factor
α_λ	Power-law exponent
$c_1 \cdots c_8$	PCA coefficients
$a_{\rm MF}$	Linear mean-flux regulation coefficient
$b_{\rm MF}$	Quadratic mean-flux regulation coefficient

quasar power-law continuum. The power-law parameter also helps account for low-order spectrophotometric errors as well as dust extinction in the spectrum.

The redshift-correction factor, $c_z = \lambda_{\rm rest}^{\rm fit}/\lambda_{\rm rest}^{\rm pipe}$, translates the spectrum along the wavelength axis with respect to the rest wavelength given by the pipeline redshift, $\lambda_{\rm rest}^{\rm pipe}$, to a best-fitting rest wavelength, $\lambda_{\rm rest}^{\rm fit}$. It is required as the SDSS pipeline redshifts are not completely accurate (see, e.g., Hewett & Wild 2010). However, due to the asymmetry and velocity shifting of quasar emission lines at different redshifts, we do not necessarily interpret c_z as a true redshift correction—it is merely an ad-hoc parameter to obtain the best-possible fit to the spectrum. The full list of free parameters for our continuum-fitting procedure is shown in Table 1 ($a_{\rm MF}$ and $b_{\rm MF}$ are free parameters for the mean-flux regulation step, described in Section 3.2).

We are now in a position to carry out the fitting procedure. First, the quasar spectrum is shifted to the quasar rest frame using the pipeline redshift and normalized at $\lambda_{\rm rest} = 1275$–$1285$ Å. We then use the least-squares fitting routine MPFIT (Markwardt 2009) to find the best-fitting set of parameters, $[c_z, \alpha_\lambda, c_j]$, given the spectrum and its noise. The initial fits from this procedure is represented by the black dashed lines in the examples shown in Figure 3.

However, while the intrinsic quasar spectrum is generally well defined redward of Lyα in the SDSS spectra, in many cases intervening metal absorption lines can be seen in the spectrum in the $\lambda_{\rm rest} \approx 1216$–$1600$ Å wavelength. To prevent these absorption features from biasing the PCA fitting, we carry out a simple iterative procedure to mask these absorption lines: using the continuum, $C_{\rm init}$, obtained from the initial least-squares fit, we mask pixels in which $f(\lambda) - C_{\rm init}(\lambda) < -2.5\,\sigma(\lambda)$, where $f(\lambda)$ and $\sigma(\lambda)$ are the observed spectrum and pipeline noise, respectively. We then make a new PCA fit redward of the Lyα emission line and repeat this process until the fit converges. In Figure 3, the final PCA fits, $C_{\rm PCA}$, are shown as black solid lines while the masked pixels are denoted by crosses.

From Figure 3, we see that the least-squares PCA fitting procedure generally works well, even with noisy (S/N $\sim$ few) spectra. The fit to the Lyα λ1216 emission line is sometimes imperfect, but unsurprising considering that the fitted range ($\lambda_{\rm rest} = 1216$–$1600$ Å) only takes partial account of the line. Comparing the initial (dashed line) and final (solid line) fits in Figure 3, we see that the red-side metal absorption lines usually have little effect on the fits, but in certain cases (e.g., Figure 3(b) and (c)) absorption line masking noticeably improves the fit.

We use the absolute flux error to quantify the goodness of the PCA fits redward of Lyα:

$$|\delta F| = \frac{\int_{\lambda_{\rm min}}^{\lambda_{\rm max}} \left| \frac{C_{\rm PCA}(\lambda) - \tilde{f}(\lambda)}{\tilde{f}(\lambda)} \right| d\lambda}{\int_{\lambda_{\rm min}}^{\lambda_{\rm max}} d\lambda}, \quad (4)$$

where $C_{\rm PCA}(\lambda)$ is the fitted continuum and $\tilde{f}(\lambda)$ is the observed spectrum smoothed by a 15 pixel boxcar to avoid biasing $|\delta F|$ in noisy spectra. $\lambda_{\rm max} = 1600$ Å and $\lambda_{\rm min} = 1225$ Å represent the range over which we calculate $|\delta F|$.

The distribution of $|\delta F|$ in the fitted data is shown in Figure 5, which plots $|\delta F|$ against the red-side signal-to-noise per pixel, $S/N_{\rm red}$, for PCA fits to a subset of the SDSS spectra as well as the mock spectra described in Section 4.

Figure 5 provides a useful diagnostic for the quality of the PCA fits on the SDSS spectra. We expect the fits to the mock spectra (green crosses) to represent the case in which the PCA eigenspectra describe the spectra nearly perfectly (see Section 4), therefore they typically have smaller values of $|\delta F|$ than the real spectra. Clearly, the SDSS spectra with large $|\delta F|$ are most likely bad fits, but note that the presence of metal absorption lines and other artifacts in the real data can bias $|\delta F|$ to larger values even for good fits (we did not mask any lines when calculating $|\delta F|$, as our metal-masking algorithm is rudimentary and sometimes masks legitimate pixels).

In practice, we carry out the PCA fitting procedure using the two different PCA templates described in Section 2.2, then for each SDSS spectrum we select the fit which gives the lower value of $|\delta F|$. Fits with $|\delta F|$ values under the 95th percentile of those from the mock spectra (red line in Figure 5) are then automatically considered good fits, while the rest are visually inspected and flagged for goodness of fit on the red side of the spectrum—approximately 90% of the spectra were adequately fit. The spectra which are not well fitted by our procedure consist mostly of objects which have strong absorption systems at $\lambda_{\rm rest} > 1216$ Å, such as metal absorption from DLAs and weak BAL quasars—while our procedure includes a simple procedure to mask metals, this does not effectively deal with very strong metal absorbers. An example of this is shown in Figure 4(a). There are also quasars with unusual spectral shapes which are not represented in the template spectra described in Section 2.2, such as quasars with weak emission lines (Figure 4(b)).

The spectra now have been had PCA fits carried out on them redward of Lyα, but the predicted continuum, $C_{\rm PCA}$, extends blueward of Lyα ($\lambda_{\rm rest} < 1216$ Å). For the objects that are well fitted on the red side of the spectrum, we expect the predicted continua to provide a reasonable prediction for the shape of the Lyα continuum blueward of Lyα, but the overall amplitude is uncertain due to the EUV–NUV power-law break described in the introduction. We now turn to the next fitting step, mean-flux regulation, to constrain the continuum amplitude.

3.2. Mean-flux Regulation

In the least-squares PCA fitting step described in the previous section, we have hitherto used no information blueward of the quasar Lyα emission line due to the absorption from the Lyα forest. However, since the absorption redshift at any point in the Lyα forest is known ($z_{\rm abs} = \lambda_{\rm obs}/1216$ Å $- 1$), the average absorption averaged over each sightline can be used to constrain the predicted continuum. In this section, we will describe the use of the mean-flux evolution of the Lyα forest, $\langle F \rangle(z)$, to regulate the amplitude and slope of the predicted PCA continuum. We refer to this as the "mean-flux regulation" step.

Using the PCA continuum $C_{\rm PCA}(\lambda_{\rm rest})$ fitted to the observed spectrum, we first extract the Lyα forest transmission $F^{\rm init}(\lambda_{\rm rest}) = f(\lambda)/C_{\rm PCA}(\lambda_{\rm rest})$ in the range $\lambda_{\rm rest} = 1041$–$1185$ Å. The extracted Lyα forest is then divided into bins, and the mean flux, $\bar{F}_{\rm bin}^{\rm init}(\lambda_{\rm bin})$, is evaluated for each bin,

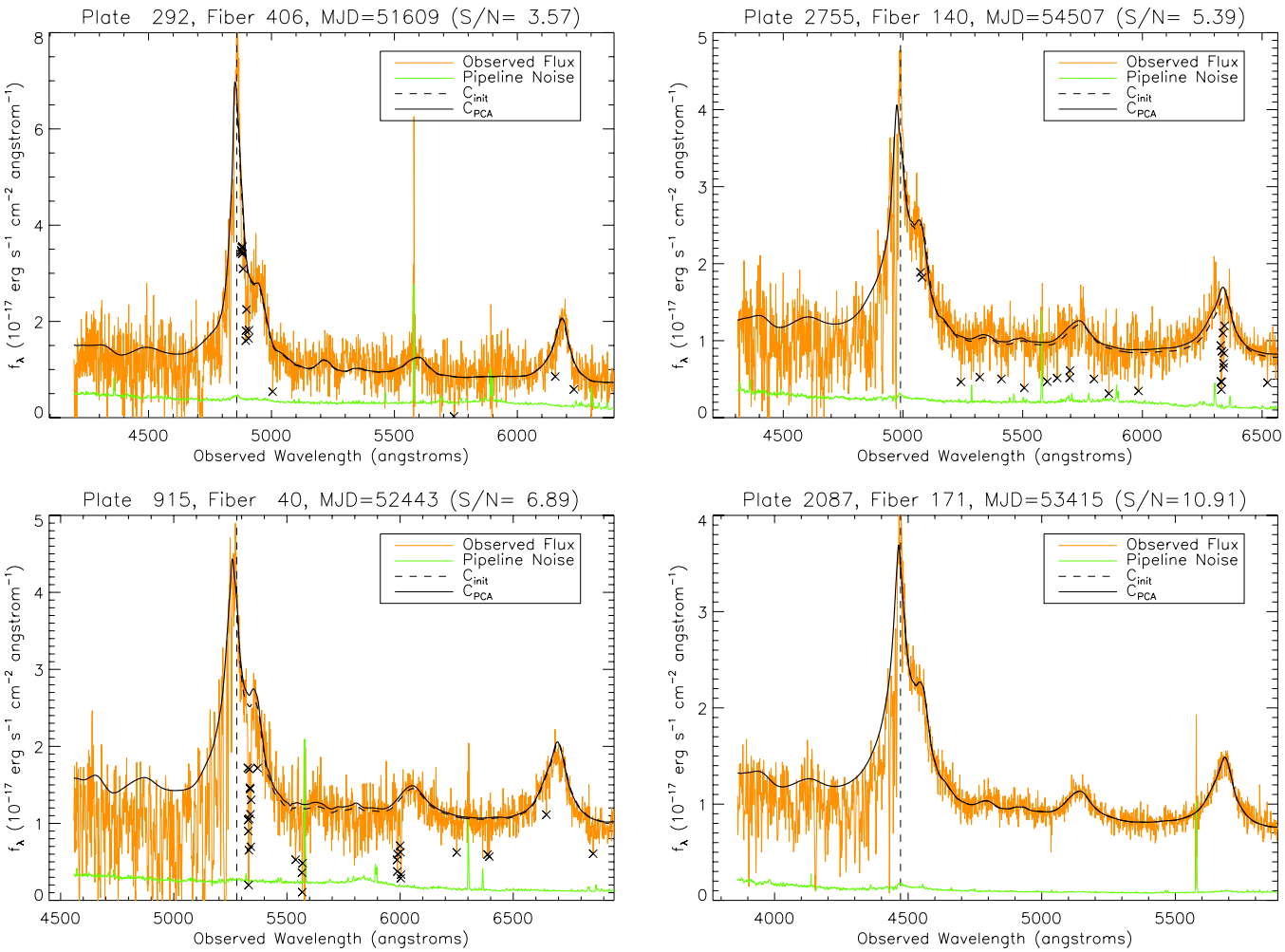

Figure 3. Successful examples of our least-squares PCA fitting method on SDSS quasar spectra with different S/N. In each plot, we show the observed flux (orange), pipeline noise (green), $C_{\rm init}$, the PCA fit from the first line-masking iteration (black dashed line), and final PCA fit, $C_{\rm PCA}$ (black solid line). Crosses indicate pixels that have been discarded by our absorption line-masking scheme. The vertical dashed lines indicate $\lambda_{\rm rest} = 1216$ Å in the quasar rest frame; all fitting is carried out redward of this wavelength. The median S/N value quoted is evaluated redward of the Lyα emission line, and the absolute flux error, $|\delta F|$, is defined in Equation (4). Note that the amplitude of the Lyα forest continuum ($\lambda_{\rm rest} < 1216$ Å) is not well fitted by the PCA procedure and will need to be corrected in the mean-flux regulation step (Section 3.2).

(A color version of this figure is available in the online journal.)

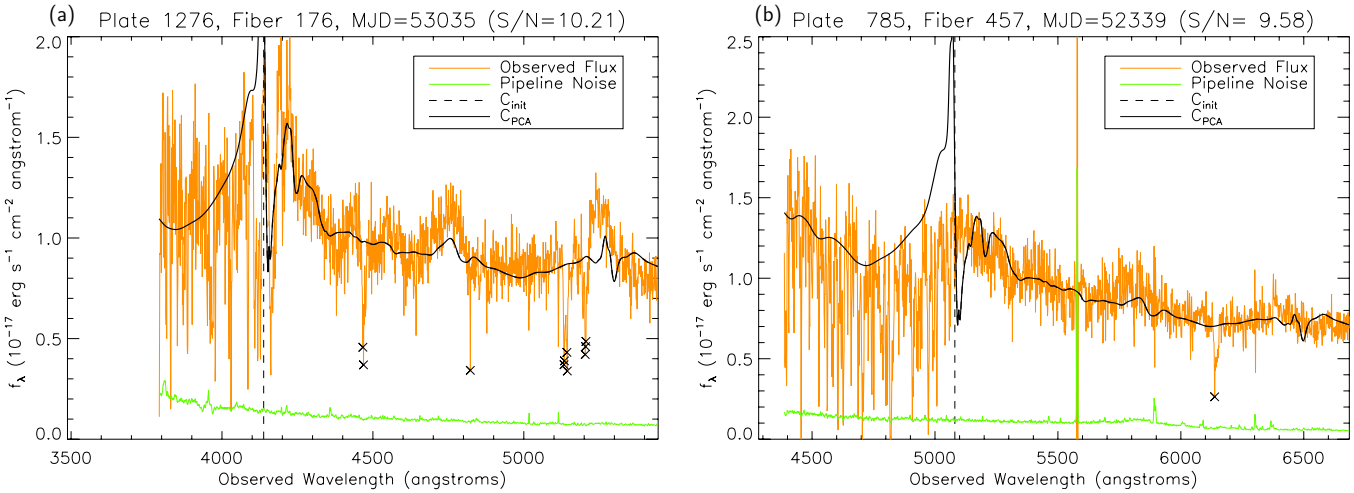

Figure 4. Examples of SDSS quasar spectra in which the PCA fitting procedure fails to provide a reasonable fit redward of the quasar Lyα line. (a) A strong proximate Lyα absorber has decimated the quasar Lyα emission line, and its associated N V+C IV have introduced broad absorption features to the $\lambda_{\rm rest} > 1216$ Å fitting region—the current algorithm is incapable of masking such strong absorption features. (b) A weak emission-line quasar. The quasar templates described in Section 2.2 do not include spectral shapes such as these.

(A color version of this figure is available in the online journal.)

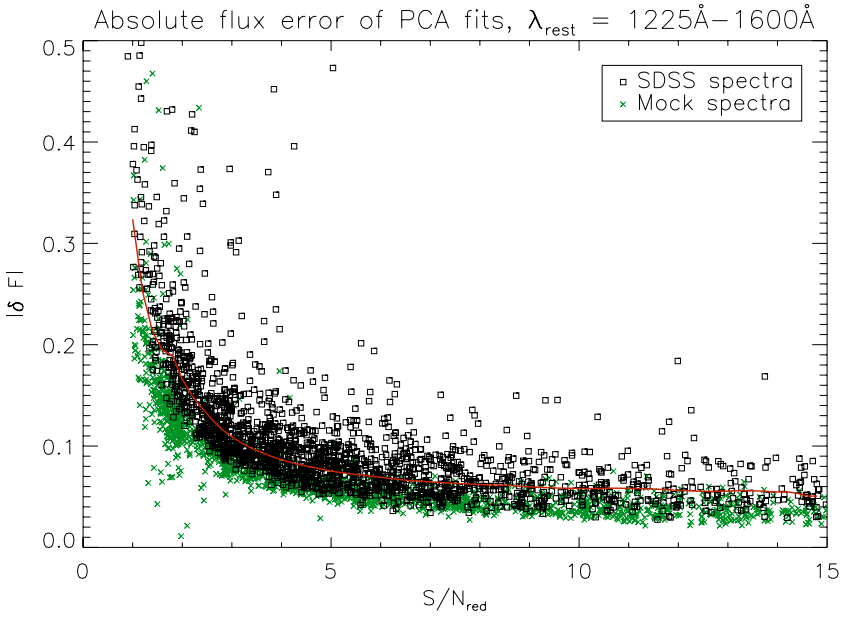

Figure 5. Dependence of absolute flux error from the red-side PCA fits, $|\delta F|$, against the S/N per pixel in the range $\lambda_{\rm rest} = 1225$–$1600$ Å. This is plotted for random subsets of 2000 SDSS spectra (black squares) and 2000 mock spectra described in Section 4 (green crosses). The red line traces the 95th percentile of $|\delta F|$ in the mock spectra; SDSS spectra with $|\delta F|$ smaller than this are automatically considered good fits, while spectra above this line are visually inspected to ensure the fit quality.
(A color version of this figure is available in the online journal.)

where $\lambda_{\rm bin} = [1070\ \text{Å}, 1110\ \text{Å}, 1150\ \text{Å}]$ are the central rest wavelengths of each bin.

We now introduce a quadratic fitting function blueward of a pivot point, $\lambda_{\rm rest} = 1280$ Å, to obtain the mean-flux-regulated continuum:

$$C_{\rm MF}(\lambda_{\rm rest}) = C_{\rm PCA}(\lambda_{\rm rest}) \times \left(1 + a_{\rm MF}\hat{\lambda}_{\rm rest} + b_{\rm MF}\hat{\lambda}^2_{\rm rest}\right), \quad (5)$$

where $a_{\rm MF}$ and $b_{\rm MF}$ are free parameters for the fit, while $\hat{\lambda}_{\rm rest} \equiv \lambda_{\rm rest}/1280\ \text{Å} - 1$. Note that for the lower redshifts ($z_{\rm QSO} \lesssim 2.4$) in which only a limited portion of the Lyα forest is accessible, we use only the linear parameter, $a_{\rm MF}$, in order to avoid overfitting.

We again use least-squares fitting to find the values of $a_{\rm MF}$ and $b_{\rm MF}$ that provide the best fit between the extracted mean-flux $\bar{F}^{\rm fit}_{\rm bin}(\lambda_{\rm bin})$ and the external mean-flux constraint $\langle F\rangle(z)$. In this mean-flux regulation step, the parameters c_z, α_λ, and c_j fitted to $\lambda_{\rm rest} > 1216$ Å are kept fixed. For the mean-flux constraint, we use the power-law-only fit from Faucher-Giguère et al. (2008) without metal correction:

$$\langle F\rangle(z) = \exp[-0.001845(1+z)^{3.924}]. \quad (6)$$

Note that we used their measurement with metals included, because we are applying these constraints to the SDSS Lyα forest spectra that have not had metals removed from their sightlines (and would probably be impossible to remove, in the case of the noisier spectra).

In principle, the errors in the continuum fit should now be at the level of a few percent, arising from some combination of the large-scale variance in the Lyα forest and errors in the fitting. In the next section, we will use mock spectra to quantify the level of continuum errors in the MF-PCA technique.

4. TESTS ON MOCK SPECTRA

The errors in the MF-PCA continuum-fitting technique can be tested by carrying the above procedure on noisy mock spectra and comparing the fitted continuum with the "true" continuum, which is known by construction. This testing process is also useful for checking our algorithm for bugs and efficiency. In this section, we will describe the process of generating realistic mock spectra and the quantitative results of the MF-PCA technique. We will also make a comparison between MF-PCA and the common technique of using the mean quasar continuum as the Lyα forest continuum.

4.1. Generating Mock Spectra

The first step is to create synthetic quasar spectra from PCA eigenspectra by making Gaussian realizations of the PCA weights, c_j, (see Equation (3)) in the manner described in Suzuki (2006). Note that this is an approximation, as the distribution of the weights may not be fully Gaussian, but it does generate realistic-looking quasar spectra. In principle, the PCA eigenspectra used to generate the mock spectra and those used in the fitting procedure should be separate but drawn from the same distribution. While we do have two sets of PCA eigenspectra (Section 2.2), they represent quasars with different luminosities. Hence, we do not expect to be able to use eigenspectra from Pâris et al. (2011) to fit mock spectra generated from the Suzuki et al. (2005) eigenspectra and vice versa. However, since we only use eight principal components in our PCA fitting step (Section 3.1), we can generate our mock spectra using 10 principal components in order to increase the uncertainty in the fitting. Nevertheless, the tests described in this section will primarily apply to the limit in which the PCA eigenspectra are a good representation of the fitted quasars, which we have argued (Section 2.2) is a reasonable assumption.

Therefore, we generate mock quasar spectra in the spectral range $\lambda_{\rm rest} = 1020$–$1600$ Å, using 10 principal components from the Suzuki et al. (2005) eigenspectra. Next, we need to introduce

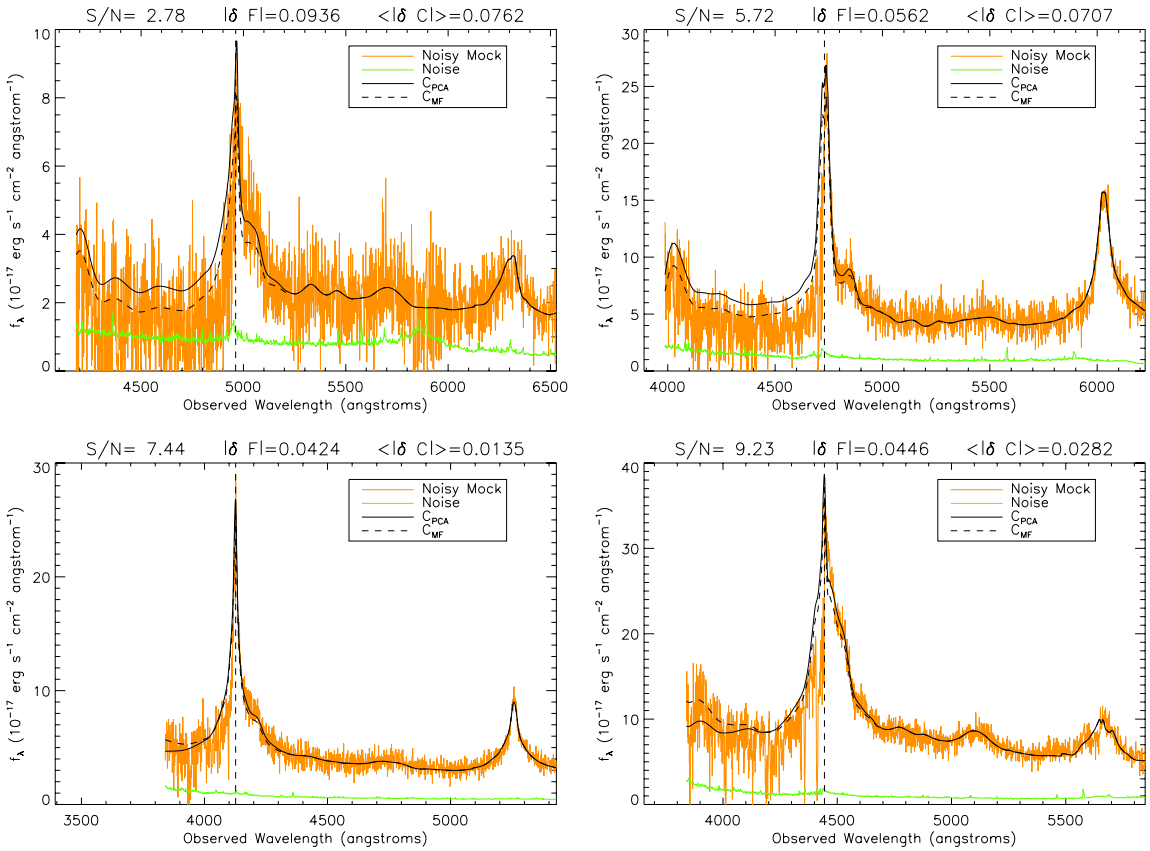

Figure 6. Tests of the MF-PCA continuum prediction procedure on mock spectra seeded with noise from real SDSS spectra. In each plot, we show the noisy mock spectrum (orange); pipeline noise (green) used to generate the mock spectrum; $C_{\rm PCA}$, the least-squares PCA fit to $\lambda_{\rm rest} = 1216$–$1600$ Å (black solid line); and $C_{\rm MF}$, mean-flux-regulated continuum fit at $\lambda_{\rm rest} < 1280$ Å (black dashed line). The vertical dashed line indicates $\lambda_{\rm rest} = 1216$ Å in the quasar rest frame. $\langle|\delta C|\rangle$ is the rms continuum-fitting error evaluated over the Lyα forest for each individual spectrum.

(A color version of this figure is available in the online journal.)

the Lyα forest absorption to the mock spectra. For this, we use the publicly available Roadrunner Lyα forest simulations[6] of White et al. (2010). These are N-body simulations with a box size of $(750\,h^{-1}\,{\rm Mpc})^3$ and a grid scale of $187.5\,h^{-1}$ kpc in which the Lyα forest flux was derived using the fluctuating Gunn–Peterson approximation. The simulations were released in the form of 22,500 Lyα forest sightlines per box, output at redshifts $z_{\rm box} \approx 2.00, 2.25, 2.50,$ and 2.75.

For a given mock quasar at redshift $z_{\rm QSO}$, we select the simulation box with the closest redshift, $z_{\rm box}$. Using Equation (6), we then re-normalize the mean flux of the box to $\langle F\rangle(z = (1+z_{\rm QSO})1100/1216 - 1)$, i.e., using the absorber redshift corresponding to the $\lambda_{\rm rest} = 1100$ Å in the quasar spectrum. We choose to normalize the mean flux across the entire box rather than in individual spectra in order to preserve the variance across different lines of sight, which is a source of error in the MF-PCA continuum fitting. A random line of sight is selected from the set of skewers, and the transmitted flux in each pixel, F_i, is rescaled to $F_i' = F_i \times \langle F\rangle(z_{{\rm abs},i})/\langle F\rangle(z = (1+z_{\rm QSO})1100/1216 - 1)$, where $z_{{\rm abs},i}$ is the absorber redshift corresponding to the pixel. This introduces redshift evolution of the mean flux, $\langle F\rangle(z)$, within the individual sightlines, which had hitherto had a fixed value of $\langle F\rangle$.

The simulated Lyα forest absorption is added to the mock quasar spectrum in $\lambda_{\rm rest} < 1216$ Å and smoothed to the approximate SDSS resolution, $R = 2000$. Gaussian noise is then added to the mock spectrum using the noise array of a randomly chosen SDSS quasar spectrum with the same $z_{\rm QSO}$ and S/N. The mock spectra are then run through the MF-PCA fitting process described above to obtain continuum fits.

4.2. MF-PCA Continuum Fitting on Mock Spectra

In Figure 6, we show several examples of the mock spectra and the fitted MF-PCA continua. The first thing to note is that mock spectra look realistic. They look similar to the real spectra shown in Figure 3, apart from the lack of metal absorption redward of the Lyα emission line.

As described in Section 3.1, we use the mock spectra as a benchmark for the PCA fit quality on the red side ($\lambda_{\rm rest} > 1216$ Å) of the spectra—we automatically accept all fits with absolute flux error, $|\delta F|$, less than the 95th percentile of the $|\delta F|$ distribution measured from the mocks (Figure 5). For the fits with larger $|\delta F|$ that require visual inspection, we use the mocks as a visual guide for what constitutes a good fit.

It is clear from Figure 6 that the mean-flux-regulated continuum, $C_{\rm MF}$, is a corrected version of the PCA fit, $C_{\rm PCA}$. In several cases, the initial PCA continuum,$C_{\rm PCA}$, appeared unphysical (e.g., dipping below the peaks of the forest). These were rectified by the mean-flux-regulated fit, $C_{\rm MF}$.

We can place this on a more quantitative footing by comparing the fitted continua, $C_{\rm fit}$, to the "true" continua, $C_{\rm true}$, which is known by construction in the mock spectra. We define the

[6] http://mwhite.berkeley.edu/BOSS/LyA/RoadRunner/

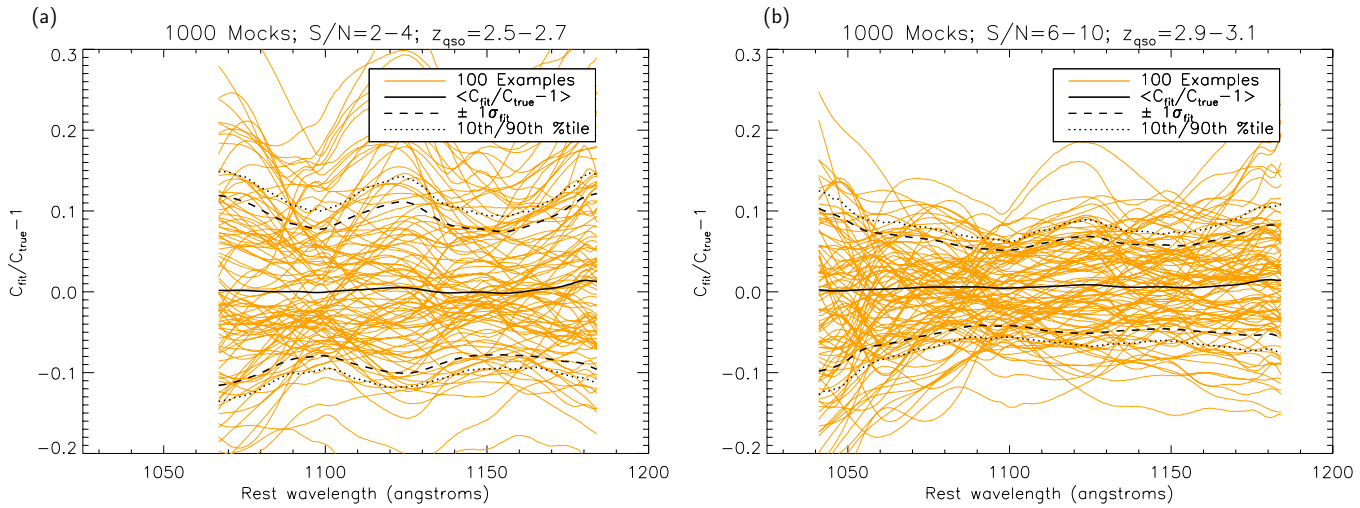

Figure 7. Continuum-fitting errors from MF-PCA fitting on mock Lyα forest spectra, plotted as a function of quasar rest-frame wavelength for (a) $2 \leqslant \mathrm{S/N} < 4$ and $2.5 \leqslant z_{\mathrm{QSO}} \leqslant 2.7$ and (b) $6 \leqslant \mathrm{S/N} < 10$ and $2.9 \leqslant z_{\mathrm{QSO}} \leqslant 3.1$. In both cases, 1000 mock spectra were generated and continuum-fitted. The gray lines represent a random subset of continuum-fitting errors from 100 fits. We also show the overall bias (solid line), the dispersion (dashed line), and the 10th/90th percentiles (dotted line) of the errors as a function of wavelength. Note that (a) represents a regime with the worst MF-PCA continua and is truncated at the lower limit of the observed spectral range, $\lambda_{\mathrm{obs}} = 3840$ Å.

(A color version of this figure is available in the online journal.)

continuum-fitting residual,

$$\delta C(\lambda_{\mathrm{rest}}) \equiv \frac{C_{\mathrm{fit}}(\lambda_{\mathrm{rest}})}{C_{\mathrm{true}}(\lambda_{\mathrm{rest}})} - 1. \quad (7)$$

We can then carry out MF-PCA fits on large numbers of mock spectra to obtain statistics on the continuum-fitting errors as a function of S/N and quasar redshift. In Figure 7, we show the residuals from fitting 1000 mock spectra in two bins of S/N and quasar redshift, z_{QSO}. The orange lines show the residuals as a function of wavelength, $\delta C(\lambda_{\mathrm{rest}})$, binned into rest frame 1 Å bins from a subset of 100 mocks. The dashed lines show the 1σ dispersion of the residuals estimated from bootstrap resampling at each 1 Å wavelength bin, while the dotted lines show the 10th and 90th percentile at each wavelength. Figure 7(a) represents one of the worst case scenarios—the signal-to-noise (S/N = 2–4) is low at $\lambda_{\mathrm{rest}} > 1216$ Å, making it difficult to obtain a good fit for the continuum shape. At the same time, the redshift is sufficiently low that the Lyα forest occupies the blue end $\lambda_{\mathrm{obs}} \lesssim 4000$ Å of the SDSS spectrographs where the S/N deteriorates rapidly with decreasing wavelength. This causes a large scatter in the flux of Lyα forest even when averaged over large segments, introducing more errors to the mean-flux regulation process.

At moderate signal-to-noise (S/N = 6–10; Figure 7(b)), the situation is significantly better. The 1σ dispersion of the fit residuals are well under 10% and 6%–7% in the central portion of the fitted region. Many of the residuals are flat to within a few percent across the Lyα forest region, indicating that the PCA-fitting has successfully accounted for the shape of the Lyα continuum. About one-tenth of the fitted continua are badly fit, with badly predicted continuum shapes and/or residuals greater than 10%.

We also calculate the mean bias, $\overline{\delta C}(\lambda_{\mathrm{rest}})$, of the residuals in Figure 7. Averaged over $\sim 10^3$ spectra, the MF-PCA technique yields a low bias of <1%, although this does not include any systematic errors in the $\langle F \rangle(z)$ measurement used to constrain the overall continuum levels. Note that this can be significant

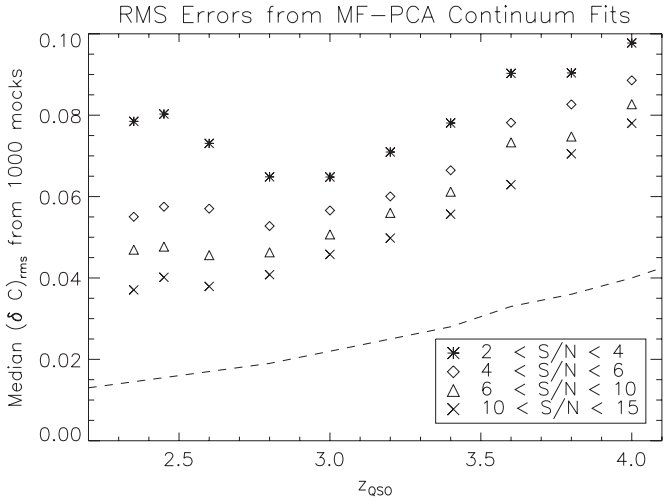

Figure 8. Median rms continuum-fitting error, $\langle |\delta C| \rangle$, from fitting 1000 mock spectra, calculated for different redshifts, z_{QSO}, and signal-to-noise ratio, S/N, on the red side of the spectrum. The rise in the low-S/N values of $\langle |\delta C| \rangle$ at low-z_{QSO} is due to the increase in noise levels at the blue end ($\lambda_{\mathrm{obs}} \lesssim 4000$ Å) of the SDSS spectra, while the overall increase in $\langle |\delta C| \rangle$ with redshift is due to the increase in the variance of the Lyα forest. The dashed line shows the rms fitting error in the absence of continuum structure and noise.

at $\langle z \rangle \gtrsim 3$, where the uncertainties in $\langle F \rangle(z)$ are at the several percent level, and increasing rapidly beyond $\langle z \rangle \sim 4$.

To quantify the overall fit quality on each mock spectrum, we use the rms of the continuum residuals evaluated over $\lambda_{\mathrm{rest}} = 1041$–$1185$ Å:

$$(\delta C)_{\mathrm{rms}} \equiv \left[\frac{\int \left(\frac{C_{\mathrm{fit}}(\lambda_{\mathrm{rest}})}{C_{\mathrm{true}}(\lambda_{\mathrm{rest}})} - 1 \right)^2 d\lambda_{\mathrm{rest}}}{\int d\lambda_{\mathrm{rest}}} \right]^{1/2}. \quad (8)$$

Figure 8 shows the median rms error from runs of 1000 mocks as a function of redshift, for four different S/N bins. At lower redshifts, the rms error is relatively high for the low-S/N spectra because the mean-flux regulation is affected by the

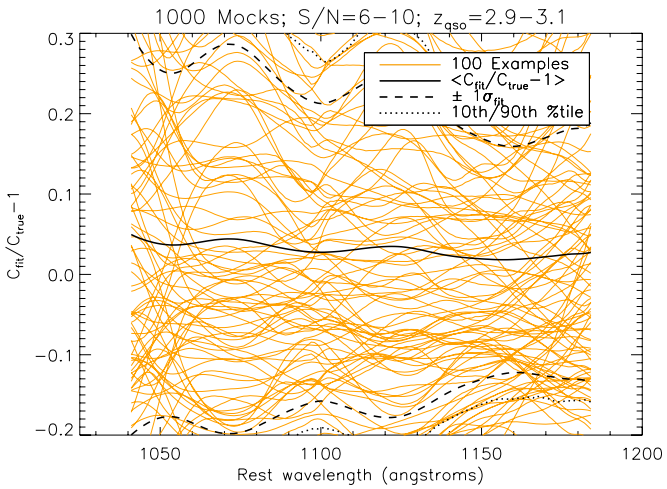

Figure 9. Same as Figure 7(b), but for the power-law+mean spectrum continua, $C_{\rm mean}$ (Equation (9)), fit to 1000 mock spectra. Note the larger errors in comparison with the MF-PCA residuals.

(A color version of this figure is available in the online journal.)

increased noise levels near the blue end of the SDSS spectra at $\lambda_{\rm obs} \sim 4000$ Å. As the observed Lyα forest region clears the blue end of the spectra, the rms error decreases to a minimum at $z_{\rm QSO} \approx 3.0$. It then rises with redshift at $z_{\rm QSO} > 3$ due to the increasing variance in the Lyα forest, which adds to the error in the mean-flux regulation. At fixed redshift, the median rms decreases with S/N as might be expected. Below $z_{\rm QSO} \approx 3$, it drops below 5% rms for moderate signal-to-noise (S/N $\sim$ 5) and asymptotes to ~3%–4% rms for the S/N > 10 spectra.

To estimate the contribution from various sources of error in the MF-PCA fitting, we also carried out the mean-flux regulation directly on the simulated Lyα forest skewers without introducing a quasar continuum, or adding pixel noise. In other words, we directly fit the function Equation (5) to the mean flux evaluated over three bins in each skewer. The rms error in the continuum from this estimate is shown as the dashed line in Figure 8. For $z_{\rm QSO} = 2.3$ quasars, the rms error contribution from the Lyα forest variance is about 1.5%; this increases to 4% at $z_{\rm QSO} = 4$. This suggests that even in the limit of high-S/N, errors in the continuum shape from PCA fitting contributes 1%–2% to the overall rms continuum error.

4.3. Power-law+Mean Continuum Fitting

To place the above results in context, we carry out another set of continuum fits on the mock spectra using the continuum model

$$C_{\rm mean}(\lambda_{\rm rest}) = f_{1280} \times \mu(\lambda_{\rm rest}) \times \left(\frac{\lambda_{\rm rest}}{1280\ \text{Å}}\right)^{-\alpha} \quad (9)$$

in which the mean spectrum, $\mu(\lambda_{\rm rest})$, is multiplied with a power law, $\lambda_{\rm rest} \propto \lambda_{\rm rest}^{-\alpha}$, and f_{1280} is the flux normalization at $\lambda_{\rm rest} = 1280$ Å. Both f_{1280} and α are determined separately for each quasar, with the power law fitted to the regions near $\lambda_{\rm rest} = 1280$ Å and $\lambda_{\rm rest} = 1450$ Å. This model is highly similar to that implemented in Slosar et al. (2011).

In Figure 9, we show the continuum residuals from $C_{\rm mean}$, as a function of rest wavelength in the Lyα forest region. Comparing this plot with Figure 7(b), which show the MF-PCA fitting residuals from the same [S/N, $z_{\rm QSO}$] bin, it is clear that MF-PCA dramatically reduces the range of fitting errors by a factor of three. Indeed, the $C_{\rm mean}(\lambda_{\rm rest})$ residuals are significantly larger than even the worst-case scenario for MF-PCA continua, represented in Figure 7(a). The significant bias (black line) of the residuals from $C_{\rm mean}(\lambda_{\rm rest})$ is puzzling at first glance, as one would expect to recover the mean spectrum (and hence no bias) when averaging over large numbers of spectra. We suspect this bias most likely due to an asymmetry in the distribution of power-law spectral indices in quasars (see Desjacques et al. 2007), which we have not accounted for in our mock spectra. However, this does not affect the scatter and shape of the continua, which is the present quantity of interest. The median rms error from the continua shown in Figure 9 is $(\delta C)_{\rm rms} = 0.13$, which is worse than any of the [S/N, $z_{\rm QSO}$] bins evaluated in Figure 8.

4.4. Residual Continuum Power

The rms continuum-fitting error is not the most important quantity for studies of the one-dimensional Lyα forest flux power spectrum:

$$P_F(k) = 2\pi \int_0^\infty \delta(k)\delta^*(k)dk, \quad (10)$$

where $\delta \equiv F/\langle F\rangle - 1$, and $k \equiv 2\pi/l$ is the Fourier wavenumber. Rather, it is the Fourier power from the continuum errors that is the troublesome systematic. For example, in their measurement of $P_F(k)$ from SDSS data, McDonald et al. (2006) were limited to scales of $k \geqslant 0.0014\,{\rm km}^{-1}$ s, corresponding to comoving distances of $r \lesssim 50\,h^{-1}$ Mpc or $\lambda_{\rm rest} \lesssim 18$ Å in the quasar rest-frame wavelength. This was due to the increasing influence of continuum power at large scales. It is therefore pertinent to investigate the amount of residual Fourier power introduced by the various continuum-fitting methods.

In Figure 10, we show the mean power spectrum of the continuum residuals from 1000 mock spectra, $\delta C(\lambda_{\rm rest})$, for the power-law+mean continuum method, $C_{\rm mean}$, and MF-PCA continuum fitting, $C_{\rm MF}$, shown for two signal-to-noise bins. All the mock spectra had quasar redshifts in the range $z_{\rm QSO} = 2.9$–3.1. All the residual power spectra have the same overall shape, with a bump at $k \approx 0.0005\,{\rm km}^{-1}$ s corresponding to the weak emission lines in the intrinsic quasar spectrum at scales of $\lambda_{\rm rest} \approx 50$ Å. This is unsurprising, since imperfections in the continuum fitting should give similarly shaped residuals. At fixed k, the amplitudes follow the same pattern we had already discussed above: the residual power from MF-PCA fitting is significantly lower than that from the power-law+mean continuum fits. Even with noisy S/N $\approx$ 3 spectra, $C_{\rm MF}$ continuum fitting reduces the residual power by ~30% compared to $C_{\rm mean}$ fitted to higher-S/N spectra.

In their measurement of $P_F(k)$ from SDSS data, McDonald et al. (2006) had used a mean continuum shape for all their spectra, which is similar to $C_{\rm mean}$ except that they did not fit for the individual power-law indices. Using this method, McDonald et al. (2006) found that residual continuum power started becoming problematic at scales greater than $k = 0.0014\,{\rm km}^{-1}$ s, indicated by the solid red vertical line in Figure 10. We thus estimate the residual power in $C_{\rm mean}$ at this scale at which continuum power interferes with $P_F(k)$ measurements. We can then look for the points in the $C_{\rm MF}$ residual power spectra with the same limiting power (red arrows in Figure 10). Note that this assumes that the Lyα forest power is constant whereas the Lyα forest power decreases with scale, but the change is gradual. For our rough estimates, we can approximate it as constant over small logarithmic intervals.

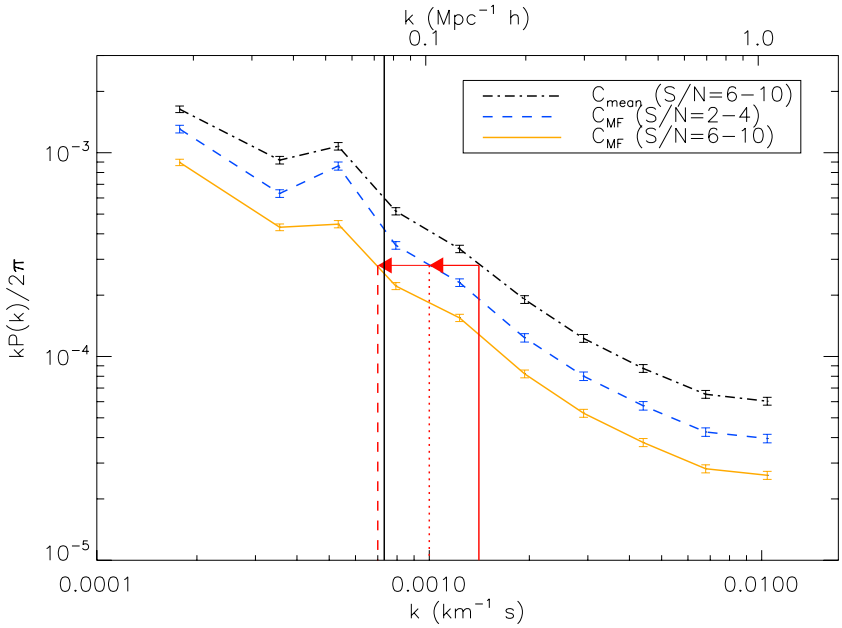

Figure 10. Mean power spectrum of 1000 residuals from continuum fitting on mock spectra, calculated using power-law+mean continuum fitting, $C_{\rm mean}$, on S/N = 6–10 spectra (black dot-dashed lines) and mean-flux-regulated PCA fitting, $C_{\rm MF}$, on spectra with S/N = 2–4 (yellow dashed lines) and S/N = 6–10 (blue solid lines). The upper abscissa shows the wave number in units of comoving distance, evaluated at $z = 2.75$ and assuming a flat ΛCDM cosmology with $h = 0.7$ and $\Omega_m = 0.28$. The black vertical triple-dot-dashed line indicates the location of the first BAO peak for this cosmology. The error bars show the error on the mean estimated from bootstrap resampling. The red solid line denotes the continuum-limited scale of $k \gtrsim 0.0014\,{\rm km}^{-1}$ s, the smallest k at which McDonald et al. (2006) measured $P_F(k)$. Red arrows indicate the points at which the $C_{\rm MF}$ fits reach the same continuum-limited power as $C_{\rm mean}$. The vertical red dotted and red dashed lines indicate the new continuum-limited scale for the S/N = 2–4 and S/N = 6–10 MF-PCA fits, respectively.

(A color version of this figure is available in the online journal.)

The corresponding limiting values of k (red dotted lines) are significantly smaller than for $C_{\rm mean}$. This suggests that MF-PCA continuum fitting could allow the Lyα forest flux power spectrum to be measured at larger scales than previously possible. For S/N = 6–10 spectra, the lower k-limit is now $k = 0.0007\,{\rm km}^{-1}$ s. This corresponds to a doubling of the accessible comoving scales: $r = 2\pi/k \approx 90\,h^{-1}$ Mpc at $z = 2.75$ compared with $r \approx 45\,h^{-1}$ Mpc in the McDonald et al. (2006) study, where these distances are calculated assuming a standard flat ΛCDM cosmology with $h = 0.7$, $\Omega_m = 0.28$ and with a $w = -1$ cosmological constant. Even for noisy (S/N = 2–4) spectra, the accessible scale has been increased significantly to $r \approx 65\,h^{-1}$ Mpc.

The new continuum-limited scales approach the $\sim 100\,h^{-1}$ Mpc BAO scale at moderate signal-to-noise (S/N $\gtrsim$ 6). In Figure 10, the black vertical triple-dot-dashed line indicates the wavenumber of first BAO peak, calculated from the prescription in Eisenstein & Hu (1998). Even though future BAO measurements in the Lyα forest are expected to be carried out in three dimensions (McDonald & Eisenstein 2007), the increase in accessible modes along the lines of sight will improve the robustness and precision of the measurements.

5. RESULTS AND CONCLUSION

5.1. Public Release of Continua

We have carried out MF-PCA continuum fitting on 12,069 quasar spectra from the SDSS DR7 catalog. The continua in the spectra range 1030 Å $< \lambda_{\rm rest} <$ 1600 Å have been made publicly available and can be downloaded via anonymous FTP.[7] The IDL fitting code took ∼0.5 s per spectrum (including file input/output) on a single processor core of a 3.0 GHz Intel Quad Core desktop with 2 GB of RAM, allowing the entire SDSS DR7 Lyα forest sample to be fitted in about two hours.

Since we expect the MF-PCA technique to provide a good continuum fit only when there is a good PCA fit redward of the quasar Lyα line, the fitted spectra have been visually inspected to verify the fit quality in the $\lambda_{\rm rest} = 1216$–1600 Å wavelength region. Approximately 89% of the spectra had reasonable PCA fits, and these have been flagged as such in the publicly available continua, although we recommend that users of the continua should make their own cuts on the fit quality. Approximately 30% of the spectra were better fitted by the low-redshift Suzuki et al. (2005) templates, while the rest were better fitted by the Pâris et al. (2011) templates. This is qualitatively as expected, since there is greater overlap by the Pâris et al. (2011) templates in the luminosity distribution of the SDSS quasars.

From the mock spectra analysis of the fit quality in Section 4, we have estimated the continuum-fitting error at each pixel within the Lyα forest, as a function of quasar redshift and spectral S/N. However, the errors have significant covariances, therefore it would be too unwieldy to provide the full error estimates for each spectrum although we can provide them upon request.

It is worth emphasizing that our continuum estimation methods require a prior assumption on the mean flux of the Lyα forest in our fits (Equation (6)). This means that our continuum estimates cannot be used to measure $\langle F \rangle(z)$; we expect our continua to be useful primarily for higher-order Lyα forest flux statistics that are less sensitive to uncertainties in $\langle F \rangle(z)$, such as the flux power spectrum.

5.2. Empirical Tests of the Fit Quality

While we have studied the performance of MF-PCA continuum fitting on mock spectra in Section 4, it is difficult to

[7] ftp.astro.princeton.edu/lee/continua/

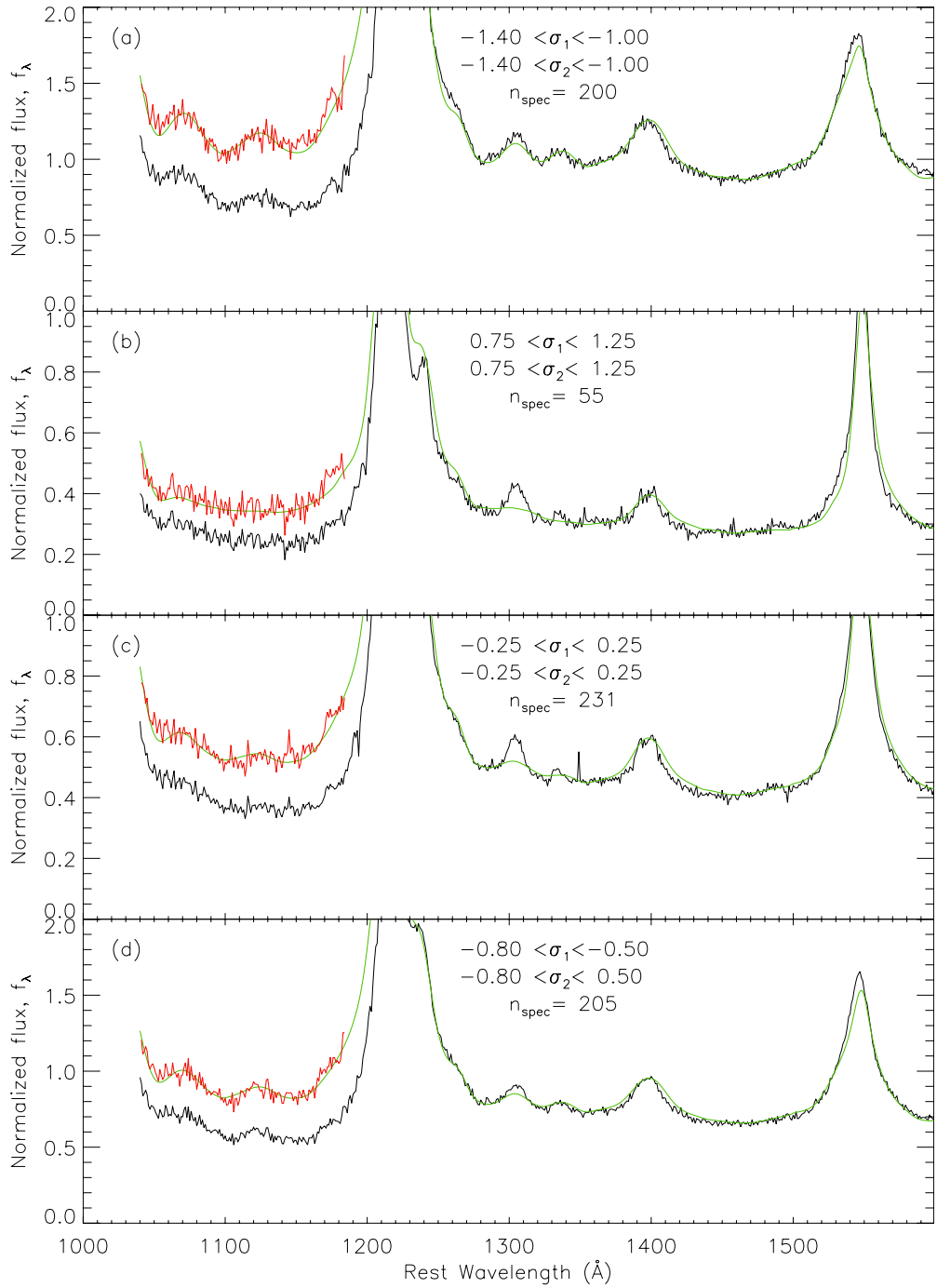

Figure 11. Stacked SDSS Lyα forest spectra (black) and similarly stacked MF-PCA continuum fits, plotted for narrow selections of the first two PCA eigenvalues, σ_1 and σ_2. The red curve shows the 1041 Å $< \lambda_{\rm rest} <$ 1185 Å Lyα forest region of the spectra which have been corrected by the mean flux prior to stacking. The agreement of the MF-PCA stacks with the mean-flux-corrected Lyα forest stacks show that the MF-PCA is doing a good job of predicting the shape of the Lyα forest continuum.

(A color version of this figure is available in the online journal.)

empirically constrain the quality of the fits. One possibility is to compare a small subset of the data with high-resolution, high-S/N spectra of the same objects. However, even with high-resolution spectra, it is questionable whether there are sufficient transmission peaks in the forest to adequately constrain the continuum shape; Faucher-Giguère et al. (2008) have shown that accurate fitting of the quasar continuum is difficult at $z \gtrsim 2.5$—even though they had corrected for the overall bias using cosmological simulations for their subsequent analysis, it is difficult to recover the true continuum shape for any single spectrum. Furthermore, most high-resolution spectra are obtained from echelle spectrographs with uncertain spectrophotometry, so it would be tricky to directly compare quasar sightlines which have been observed in both SDSS and high-resolution echelle spectrographs.

However, it is possible to get a sense of the efficacy of our MF-PCA continua by stacking large numbers of spectra. This cancels out the Lyα forest power from individual sightlines and allows the underlying continuum shape to be seen, albeit lowered due to the mean Lyα absorption.

Recall that the PCA coefficients, c_j, parameterize the shape of the quasar spectra. Therefore, if our continuum-fitting technique works, quasars with similar c_j measured from $\lambda_{\rm rest} > 1216$ Å should have similar-looking continua within the Lyα forest region. Thanks to the large number of spectra in our SDSS sample, it is possible to stack $\sim 10^2$ spectra with similar values of c_j to recover the collective shape of their Lyα forest continua.

We can select subsamples of quasars based on their values of $\sigma_1 \equiv c_1/\lambda_1$ and $\sigma_2 \equiv c_2/\lambda_2$, where $\lambda_1 = 7.563$ and $\lambda_2 = 3.604$ are the standard deviations of c_1 and c_2, respectively, in the low-redshift *HST* eigenspectra (Suzuki 2006). These two principal components account for approximately 80% of the total variance in the low-redshift quasar templates. We limit ourselves to spectra with S/N $> 3\,{\rm pixel}^{-1}$, and which have been visually inspected to be decent fits redward of $\lambda_{\rm rest} = 1216$ Å. We also select quasars with $z_{\rm QSO} > 2.6$ in order to ensure reasonably complete coverage of the Lyα forest. Within a subsample, each spectrum is first normalized near $\lambda_{\rm rest} = 1280$ Å and rebinned into a common wavelength grid with $\Delta\lambda_{\rm rest} = 1$ Å bins before being stacked. The same procedure is carried out on the MF-PCA continuum fitted to each spectrum to obtain a mean MF-PCA continuum for the subsample.

In Figure 11, we show four subsamples from our SDSS sample with different $[\sigma_1, \sigma_2]$ with respect to the low-redshift Suzuki et al. (2005) eigenspectra. Redward of 1216 Å, we see that the least-squared PCA procedure generally does a good job of fitting the emission lines, although there are inaccuracies in fitting N v λ1240 and Si ii λ1306. Blueward of 1216 Å, the stacked spectrum appears to have a similar shape to the fitted continua, although the overall flux level is depressed due to the mean Lyα absorption.

We can make a more direct comparison between the stacked spectra and the fitted Lyα forest continua by correcting each observed Lyα forest pixel by its mean flux (using Equation (6)) before stacking. The mean-flux-corrected Lyα forest is shown as the disembodied red line in Figure 11. It is gratifying to see that the stacked MF-PCA continua generally agrees well with the stacked Lyα forest spectra. Our technique can clearly account for the diversity in quasar continua: Spectra with clear emission-line features (Figure 11(a)) and those with smooth continua (Figure 11(b)) are well differentiated. Because we have corrected each Lyα forest pixel by the same mean flux (Equation (6)) that we have used to carry out the MF-PCA fits, we expect the amplitude of the corrected Lyα forest stacks, in $\lambda_{\rm rest} = 1041$–$1185$ Å, to match those of the stacked MF-PCA continua, but the tilt and shape of the continuum bears testament to the success of the technique. In addition, since the MF-PCA continua shown in Figure 11 were a subset which used the Suzuki et al. (2005) quasar templates, this suggests that it was appropriate to use low-redshift templates to fit some of the $z_{\rm QSO} \gtrsim 2$ SDSS spectra.

5.3. Conclusions

We have introduced MF-PCA continuum fitting, a new technique for predicting the Lyα forest continuum in low-S/N spectra. In tests on mock spectra, we have found that MF-PCA can predict the continuum at the 8% rms level in SDSS spectra with S/N ~ 2 at $z = 2.5$, and <5% rms in S/N $\gtrsim 5$ spectra. This is a significant improvement over the ∼15% rms continuum errors previously achievable in low-S/N spectra. We are making available MF-PCA continuum fits for 12,069 Lyα forest spectra from the SDSS DR7 quasar catalog. The MF-PCA technique also significantly reduces the Fourier power from continuum-fitting residuals by a factor of a few in comparison with dividing by a mean continuum. This will allow a concomitant increase in the accessible scales for Lyα forest flux power spectrum measurements.

This improved continuum-fitting accuracy will significantly increase the value of low-S/N Lyα forest data. For example, the ongoing Baryon Oscillations Spectroscopic Survey (BOSS; Eisenstein et al. 2011) will obtain Lyα forest spectra from ∼150,000 quasars at $z_{\rm QSO} \gtrsim 2$ (Ross et al. 2011), with the aim of measuring the BAO feature in the Lyα forest absorption across different quasar sightlines (e.g., Slosar et al. 2011). The typical signal-to-noise (S/N ∼ 2) of BOSS spectra will be even lower than that of SDSS (S/N ∼ 4), therefore we expect the MF-PCA technique to contribute significantly to the utility of the BOSS data.

There are several ways in which the current work could be improved. The Suzuki et al. (2005) PCA templates, which we have used to fit some of the spectra, which were derived from a low-redshift quasar sample may not be a perfect descriptor of the SDSS data (although the test in Section 5 shows that it does a reasonable job). In addition, while the $z_{\rm QSO} \sim 3$ Pâris et al. (2011) templates were indeed obtained from SDSS quasars, they used a high-luminosity subset that is not representative of the full SDSS luminosity distribution. Furthermore, the hand-fitting technique that they had used to obtain continua from these spectra cannot be used for the lower-luminosity (and hence lower-S/N) quasars.

However, with a large data set such as SDSS or BOSS, it is possible to regard the Lyα forest absorption within individual spectra as a noise term that cancels out with sufficiently large numbers of template spectra. This will allow new eigenspectra to be generated from the data itself, although each individual spectrum will need to be corrected by its mean flux before being included in the eigenspectrum solution. In the near future, we will work on this technique to generate new eigenspectra from the BOSS data.

The other issue with the MF-PCA fitting is that it requires an assumed mean flux, $\langle F \rangle(z)$ for the Lyα forest. This is not ideal, as the evolution of the mean flux is an important observable of the Lyα forest. This could in principle be overcome by solving simultaneously for the mean flux of the Lyα forest and the continuum-fitting parameters for the individual spectra, using maximum-likelihood techniques. This could potentially allow large Lyα forest data sets to be continuum-fitted and studied in a fully self-consistent fashion.

The authors thank Michael Strauss and Xavier Prochaska for useful discussions and comments, and to Isâbelle Paris for providing the data from her mean-flux measurement.

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web site is http://www.sdss.org/.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions: the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns

Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

REFERENCES

Allen, J. T., Hewett, P. C., Maddox, N., Richards, G. T., & Belokurov, V. 2011, MNRAS, 410, 860
Baldwin, J. A., Burke, W. L., Gaskell, C. M., & Wampler, E. J. 1978, Nature, 273, 431
Bernardi, M., Sheth, R. K., SubbaRao, M., et al. 2003, AJ, 125, 32
Bolton, J. S., Haehnelt, M. G., Viel, M., & Springel, V. 2005, MNRAS, 357, 1178
Croft, R. A. C., Weinberg, D. H., Bolte, M., et al. 2002, ApJ, 581, 20
Croft, R. A. C., Weinberg, D. H., Pettini, M., Hernquist, L., & Katz, N. 1999, ApJ, 520, 1
Dall'Aglio, A., Wisotzki, L., & Worseck, G. 2009, arXiv:0906.1484
Desjacques, V., Nusser, A., & Sheth, R. K. 2007, MNRAS, 374, 206
Draine, B. T. (ed.) 2011, Physics of the Interstellar and Intergalactic Medium (Princeton: Princeton Univ. Press)
Eisenstein, D. J., & Hu, W. 1998, ApJ, 496, 605
Eisenstein, D. J., Weinberg, D. H., Agol, E., et al. 2011, AJ, 142, 72
Fan, X. 2006, New Astron. Rev., 50, 665
Faucher-Giguère, C.-A., Lidz, A., Zaldarriaga, M., & Hernquist, L. 2009, ApJ, 703, 1416
Faucher-Giguère, C.-A., Prochaska, J. X., Lidz, A., Hernquist, L., & Zaldarriaga, M. 2008, ApJ, 681, 831
Francis, P. J., Hewett, P. C., Foltz, C. B., & Chaffee, F. H. 1992, ApJ, 398, 476
Hewett, P. C., & Wild, V. 2010, MNRAS, 405, 2302
Kirkman, D., Tytler, D., Suzuki, N., et al. 2005, MNRAS, 360, 1373
Lee, K.-G. 2011, arXiv:1103.2780
Lee, K.-G., & Spergel, D. N. 2011, ApJ, 734, 21
Mandelbaum, R., McDonald, P., Seljak, U., & Cen, R. 2003, MNRAS, 344, 776
Markwardt, C. B. 2009, in ASP Conf. Ser. 411, Astronomical Data Analysis Software and Systems XVIII, ed. D. A. Bohlender, D. Durand, & P. Dowler (San Francisco, CA: ASP), 251
McDonald, P., & Eisenstein, D. J. 2007, Phys. Rev. D, 76, 063009
McDonald, P., Miralda-Escudé, J., Rauch, M., et al. 2000, ApJ, 543, 1
McDonald, P., Seljak, U., Burles, S., et al. 2006, ApJS, 163, 80
Noterdaeme, P., Petitjean, P., Ledoux, C., & Srianand, R. 2009, A&A, 505, 1087
Pâris, I., Petitjean, P., Rollinde, E., et al. 2011, A&A, 530, A50
Richards, G. T., Fan, X., Newberg, H. J., et al. 2002, AJ, 123, 2945
Ross, N. P., Myers, A. D., Sheldon, E. S., et al. 2011, arXiv:1105.0606
Schneider, D. P., Richards, G. T., Hall, P. B., et al. 2010, AJ, 139, 2360
Shen, Y., Richards, G. T., Strauss, M. A., et al. 2011, ApJS, 194, 45
Slosar, A., Font-Ribera, A., Pieri, M. M., et al. 2011, J. Cosmol. Astropart. Phys., JCAP09(2011)001
Stoughton, C., Lupton, R. H., Bernardi, M., et al. 2002, AJ, 123, 485
Suzuki, N. 2006, ApJS, 163, 110
Suzuki, N., Tytler, D., Kirkman, D., O'Meara, J. M., & Lubin, D. 2005, ApJ, 618, 592
Telfer, R. C., Zheng, W., Kriss, G. A., & Davidsen, A. F. 2002, ApJ, 565, 773
Tytler, D., Kirkman, D., O'Meara, J. M., et al. 2004, ApJ, 617, 1
vanden Berk, D., Yip, C., Connolly, A., Jester, S., & Stoughton, C. 2004, in ASP Conf. Ser. 311, AGN Physics with the Sloan Digital Sky Survey, ed. G. T. Richards & P. B. Hall (San Francisco, CA: ASP), 21
Vanden Berk, D. E., Richards, G. T., Bauer, A., et al. 2001, AJ, 122, 549
Viel, M., Matarrese, S., Heavens, A., et al. 2004, MNRAS, 347, L26
White, M., Pope, A., Carlson, J., et al. 2010, ApJ, 713, 383
Zheng, W., Kriss, G. A., Telfer, R. C., Grimes, J. P., & Davidsen, A. F. 1997, ApJ, 475, 469