From molecular force generation Large scale structure with Lyman-[alfa] absorption surveys

Andreu Font Ribera

Large Scale Structure with Lyman-α Absorption Surveys

Andreu Font Ribera

Universitat de Barcelona

Advisor: Jordi Miralda Escudé

July 2011

Acknowledgements

To start with, I would like to express my gratitude to my advisor, Jordi Miralda. Thank you so much for your help, orientation and dedication received during these years, and for giving me the opportunity to get involved in the BOSS collaboration.

I am deeply grateful to Patrick McDonald, for his patience with my poor coding skills and for answering to my huge number of technical e-mails.

I want to thank Anze Slosar, Patrick Petitjean, Nic Ross, Jean-Marc LeGoff, Shirley Ho, David Weinberg and the rest of the Lyα forest working group in the BOSS collaboration, for attending to my talk on mock catalogs in four different meetings without complaining. See you in the next telecon.

Martin White offered me the possibility to discover Berkeley, and its incredible scientific environment, where I had the chance to discuss with Uros Seljak, Matt McQuinn, Nao Suzuki, Hee-Jong Seo and himself. I really appreciate the opportunity.

Vaishali Bhardwaj and Hector Gil helped me with the introduction of the thesis, and I am afraid I will have to pay them back soon.

Finalment, vull donar les gràcies a la meva família i a tots els que heu conviscut amb mi aquests quatre anys, ja sigui compartint pis, despatx, equips de futbol o senzillament una bona estona. Sense vosaltres tampoc hagués pogut escriure aquesta tesis.

Moltíssimes gràcies! Thank you so much!

Resum de la Tesi

Gran part dels estudis presentats en aquesta tesi doctoral estan vinculats als esforços per detectar les oscil·lacions acústiques dels barions (BAO, de l'anglès Baryon Acoustic Oscillations) en la funció de correlació de l'absorció en l'espectre de quasars llunyans, relacionada amb la transició de Lyman-α (Lyα) en els atoms d'hidrogen intergalàctic. Aquesta absorció és coneguda com a "bosc de Lyman-α".

El Baryon Oscillation Spectroscopic Survey (BOSS) és un dels quatre projectes que componen la tercera fase de l'Sloan Digital Sky Survey (SDSS-III) i té com a principal objectiu la detecció del senyal de BAO en la correlació de galàxies i en la del bosc de Lyα. Per fer-ho, des de la tardor de 2009 s'estan obtenint espectres electromagnètics de centenars de milers de galàxies i quàsars utilitzant un telescopi de 2.5 metres de diàmetre a Apache Point (Nou Mèxic, Estat Units d'Amèrica).

La meva vinculació amb BOSS començà el gener de 2009, i des d'un bon principi m'he centrat en generar simulacres d'espectres amb absorció de Lyα. Aquests simulacres han tingut una funció essencial en la primera publicació de la col·laboració (Slosar et al., 2011).

L'estructura de la tesi es divideix en quatre capítols, que resumeixo a continuació.

Simulant la Mesura de l'Espectre de Potències del Bosc de Lyα en un Catàleg Espectroscòpic de Quàsars a Grans Escales

En el capítol 2 presento un mètode per simular l'absorció de Lyα en l'espectre de quàsars. El mètode, desenvolupat en col·laboració amb el Dr. Patrick McDonald i el Dr. Jordi Miralda-Escudé, permet generar espectres amb qualsevol distribució de flux i amb qualsevol espectre de potències.

El Bosc de Lyα en Tres Dimensions: Mesura de la Correlació del Flux a Grans Escales en les Dades del Primer Any de BOSS

La primera detecció de la correlació a grans escales del bosc de Lyα va ser presentada a mitjans 2011 per la col·laboració SDSS-III, utilitzant les dades obtingudes durant el primer any del projecte BOSS (Slosar et al., 2011). El capítol 3 conté un resum de l'estudi, fent èmfasi en la meva aportació i en el paper dels simulacres explicats en el capítol 2.

L'Efecte dels Sistemes d'Alta Densitat de Columna en la Mesura de la Funció de Correlació del Bosc de Lyα

En el Capítol 4, presento un estudi analític de l'efecte que els sistemes d'alta densitat de columna tenen en la mesura de la funció de correlació del bosc de Lyα. A continuació presento un mètode per introduir aquests sistemes en els espectres simulats, desenvolupat també amb la col·laboració de Miralda-Escudé i McDonald.

Correlacions creuades del Bosc de Lyα

En el darrer capítol d'aquesta tesi, 5, presento un estudi sobre la possibilitat de detectar la correlació creuada entre una població de galàxies i l'absorció de Lyα. Presento també un mètode senzill per mesurar la correlació creuada en un catàleg espectroscòpic com ara BOSS, i estudio en concret la possibilitat de mesurar el biaix dels sistemes Lyα esmorteits ("Damped Lyα systems" en anglès).

Contents

CHAPTER 1

Introduction

1.1 Modern Cosmology

1.1.1 The Big Bang Paradigm

The basis of modern cosmology was established during the first decades of the twentieth century. Alexander Friedmann, Howard Percy Robertson and Arthur Geoffrey Walker derived a solution to Einstein's equations of General Relativity with the symmetries imposed by the Cosmological Principle, namely, the Universe is isotropic and homogeneous on large scales. The conclusion was that such a universe would experience a homogeneous contraction or expansion, with gradual deceleration due to gravitational attraction.

In 1929, Edwin Hubble published a study with the radial velocities of 46 galaxies, computed from the Doppler displacement present in their electromagnetic spectra. Hubble showed that most galaxies are receding from us, and that there is a clear correlation between distance and recession velocity. This relation, now known as Hubble law, had been predicted two years before by George Lemaître as a natural consequence of the expansion or contraction of the Universe.

Since the galaxies are receding from us in an expanding Universe, the shift in the electromagnetic spectrum is always to the red (i.e. lower energies), and is consequently referred to as redshift. The redshift, z, of an object is then a measure of its recessional velocity, and using the Hubble law, also a measure of distance. Finally, because of the finite value of the speed of light, redshift is also a measure of time.

During the following decades there were constant debates concerning two different paradigms that explained the expansion. The Big Bang model, first proposed by Lemaître and developed by George Gamow and others, claimed that the Universe was once in a very dense and hot state and has been expanding ever since, as a result of a primordial explosion (hence the name of Big Bang). The Steady State theory, developed by Fred Hoyle, Thomas Gold, Hermann Bondi among others, proposed a stationary universe where matter is continuously created as the Universe expands allowing its properties to remain constant. In this way, the steady state model avoided the need for an origin of the Universe.

One of the main virtues of the Big Bang model is that it naturally explains the abundances of the most elementary elements in the Universe (hydrogen, helium, lithium, etc.), created during what is known as Big Bang Nucleosynthesis (BBN). Another prediction, introduced by Ralph Alpher and Bob Herman, is the presence of a cosmic microwave background (CMB), an homogeneous background radiation that holds information from the earlier phases of the Universe. In 1964, when Arno Penzias and Robert Wilson serendipitously discovered the CMB, the Big Bang model obtained the status of the most accepted model for the origin of the Universe.

1.1.2 Primordial Fluctuations and Dark Matter

In 1934, Fritz Zwicky presented a study of the dynamics in the galaxy cluster of Coma. He showed that the depth of the gravitational well inferred by the velocity dispersion of galaxies could not be explained by the mass of stars present in galaxies, and he postulated the presence of Dark Matter (DM). Several decades after Zwicky, the measurement of galactic rotation curves carried out by Vera Rubin provided an independent probe of the existence of dark matter. Dark matter may be any type of matter that only interacts gravitationally or extremely weakly through any other interaction. Its presence may therefore be impossible to detect except through its gravitational influence. The Big Bang theory was soon expanded to include a dark matter component, and the theory that was best consistent with observations was named Cold Dark Matter (CDM), where cold refers to the need for a low primordial velocity dispersion of the dark matter.

In the CDM theory, the large scale structure of the present universe originates from small fluctuations in the early universe. On cosmological small scales matter is clearly not distributed homogeneously, but forms galaxies, clusters of galaxies, voids, etc. The gravitational attraction progressively increases these initial, small fluctuations, creating the different levels of structure present today. The CDM theory was able to predict the spectrum of anisotropies present in the large scale structure and in the CMB, assuming Gaussian and scale invariant initial conditions.

During the 1980s, an extension of the theory of the Big Bang was introduced by Alexei Starobinsky and Alan Guth. The theory proposed that during its very early stages, the Universe underwent an exponential and sudden phase of expansion that caused an increase of many tens of orders of magnitude of its size. This theory, coined by the name of inflation by Guth, was able to account for the primordial perturbations as quantum fluctuations in the early Universe, and gave a natural explanation for the nearly scale-invariance of the primordial power spectrum.

The inflationary CDM theory was extensively accepted by the cosmological community when the COBE satellite first, and the WMAP later,measured the power spectrum of the primordial fluctuations imprinted in the CMB.

1.1.3 Accelerated Expansion of the Universe

Using the measurement of the luminosity distance of Type Ia Supernovae, the Supernova Search Team (1998) and the Supernova Cosmology Project (1999) detected an accelerated expansion of the Universe. This result has been confirmed by different cosmological probes since. Following the laws of general relativity (GR), the accelerated expansion cannot be explained by a universe containing only matter and radiation.

The simplest explanation for the accelerated expansion is to consider a non-zero value for the integral constant in Einstein's Equations, the cosmological constant Λ. Alternative explanations include modifications of General Relativity (GR) and the presence of an extra field in the Universe with an effective negative pressure, referred to as Dark Energy (DE).

Many different observational tests have been proposed in order to shed light on the nature of the accelerated expansion of the Universe. These include methods to measure the growth rate of density perturbations with time, and the study of the geometry of the Universe and its evolution. Most of the latter studies attempt to measure the Hubble parameter $H(z)$ or related observables at different redshifts, with the maximum precision.

One of the most promising approaches is to measure the size of the Baryon Acoustic Oscillations (BAO) scale in the clustering of matter at different redshifts. The early Universe was very dense and high energy photons and ions were tightly coupled, and their interactions prevented the first atoms to form. The competition between the gravitational attraction and the baryonic pressure produced sound waves in the photon-baryon fluid. The density and temperature of the Universe dropped with the expansion and eventually the photons decoupled from the baryons, and the Universe became neutral. After this epoch, known as the "recombination epoch", the baryons were left with an imprint caused by the sound waves, with a characteristic scale set by the sound horizon at the moment of decoupling $r_{BAO} \sim 150\,\mathrm{Mpc}$. This imprint is seen in the power spectrum of CMB anisotropies at a very high redshift $z = 1100$, but also in the clustering of galaxies at a later time of $z = 0.3$ as measured by the Sloan Digital Sky Survey.

1.2 Lyman α Forest

1.2.1 Historical Overview

During the 1950s, radio telescopes detected unusual objects showing little or no optical counterparts, or very dim point sources with strange emission lines. These quasi-stellar radio objects were soon called "quasars", and their nature was gradually unveiled in the following decades.

Today, the consensus is that quasars are compact regions near the center of a galaxy surrounding its central super massive black hole. These objects are one of the brightest objects in the Universe and hence can be detected at very high redshift, i.e. at large distances and earlier times.

In 1965, Maarten Schmidt spectroscopically observed the quasar 3C 9 and reported a redshift of $z = 2.01$. In this case, the Lyα emission line present in all quasars was redshifted to the optical part of the spectrum. The non-zero flux detected at energies higher than the Lyα transition allowed James Gunn and Bruce Peterson to place a strong upper limit to the abundance of neutral hydrogen in the intergalactic medium (IGM). The effect, known as the Gunn-Peterson trough, consists of the following: photons emitted at energy above the Lyα transition will be gradually redshifted and eventually have the exact frequency to excite neutral hydrogen atoms. If a small fraction of the IGM is neutral, the Lyα absorption causes a dramatic decrease in the fraction of transmitted flux at energy above the Lyα emission line of the quasar. The fact that most of the flux is transmitted at redshifts $z = 2 - 4$, implies that the IGM is highly ionized at this redshift range.

John Bahcall and Edwin Salpeter (1965) proposed that the multiple absorption features observed in the spectra of quasars could be produced by clumps of intervening neutral hydrogen in the line of sight. This phenomenon, which became known as Lyman Alpha Forest, was confirmed by observations carried out by Roger Lynds (1971). Though early models described the absorbers as collapsed clouds with a high neutral hydrogen density, a new scenario was envisioned during the 1990s, where Lyα forest arises from the fluctuating IGM, tracing the promordial density fluctuations in the Universe (for a historical review of the Lyα forest, see Rauch, 1998).

Numerical simulations by Cen et al. (1994) helped to establish the cosmological explanation of the Lyα forest, and it was confirmed with the first measurements of its power spectrum (Croft et al., 1998, McDonald et al., 2000).

1.2.2 Lyα Forest as a Cosmological Tool

Once the cosmological origin of the Lyα absorption was established, cosmologists became interested in this new tool to study the Universe at redshifts higher than the ones probed by galaxy surveys. McDonald et al. (2006) measured, with very high accuracy, the Lyα power spectrum along the line of sight using a few thousand quasars from SDSS. This measurement put strong constraints on cosmological parameters, particularly in the neutrino masses, and motivated new efforts to design even larger quasar spectroscopic surveys to study the Lyα forest.

McDonald et al. (2006) and earlier studies only used the correlation along single lines of sight, the so-called 1D power spectrum. In order to study the correlation across different lines of sight, the density of quasars per square degree must be very large, but the advantages of doing so have become clear.

McDonald and Eisenstein (2007) computed the characteristics that a Lyα absorption survey needs in order to detect the BAO feature. The results were promising: it could be done using the infrastructure of a spectroscopic galaxy survey, causing a small overhead of ~ 20 %.

The measurement of BAO at $z = 2 - 3$ using the Lyα forest was included as one of the main goals of the Baryonic Oscillation Spectroscopic Survey (BOSS), which is the main survey of the SDSS-III collaboration. BOSS observations started in the fall of 2009 and will be completed by the spring of 2014.

1.3 Description of the Thesis

Most of the work presented in this PhD thesis is related to the efforts of detecting the BAO signature in the correlation of the Lyα forest using spectroscopic data from the BOSS survey (part of the SDSS-III collaboration, Eisenstein et al., 2011).

I joined the collaboration in early 2009, and soon became responsible to develop a code to generate mock realizations of the survey, in collaboration with Patrick McDonald and Jordi Miralda-Escudé. Mock catalogs are essential to test the data analysis code and to study the effect of possible systematics. In Chapter 2 I explain the method developed and show how it can be used to estimate uncertainties in the correlation function.

In Chapter 3 I present the first publication of the Lyα working group of BOSS collaboration (Slosar et al., 2011). In this paper, we detect for the first time the correlations of Lyα absorption on cosmologically large scales. I introduce the collaboration, give a brief summary of the publication and explain my personal work in the paper. The complete article can be found in Appendix B.

The simulated Lyα spectra were also useful to correctly interpret the results obtained Slosar et al. (2011). We showed that the presence of high column density systems in the spectra could bias the measurement of the correlation function of the Lyα forest. In Chapter 4 I explain a method to add these systems into mock spectra, developed in collaboration with Miralda-Escudé and McDonald. I also study analytically the effect of these systems in the inferred bias parameters from a spectroscopic survey, and quantify the expected effect in the measurement of Slosar et al. (2011).

Finally, in Chapter 5 I comment the possibility of detecting the crosscorrelation of galactic objects with the Lyα forest. In particular, I discuss the probability of measuring the bias of Damped Lyα systems and the mass of its host halos from the BOSS survey.

CHAPTER 2

Simulating the Lyman α Forest Power Spectrum Measurement from a Large-Scale Quasar Spectroscopic Survey

2.1 Introduction

The hydrogen Lyα absorption spectra of high-redshift sources are being revealed as an extremely powerful tool for the study of large-scale structure in observational cosmology. The numerous absorption features observed in the spectra of quasars usually described as the " Lyα forest " were originally interpreted as discrete gas clouds, but have been better understood and described as arising from the continuous cosmic web of filamentary structures that is expected in the Cold Dark Matter model of structure formation. Results from hydrodynamic cosmological simulations have shown that the observed properties of the Lyα forest are generally in good agreement with the hypothesis of a photoionized intergalactic medium with density fluctuations that are related to the same primordial perturbations that give rise to the galaxy distribution and the Cosmic Microwave Background fluctuations (e.g., McDonald et al., 2006, Rauch, 1998). The Lyα forest spectra should therefore be considered as a continuous field of the Lyα transmitted fraction $F(\mathbf{x})$ (where $\mathbf{x}$ is the redshift-space coordinate), which is related to the variations of the gas density, peculiar velocity and temperature along the line of sight, and eventually to the primordial density field, particularly on large scales, in which the complexities of non-linear evolution become less important.

In fact, if we have a large number of absorption spectra from different sources covering a large volume and with a sufficiently dense sampling, one can measure the redshift space power spectrum of the field $F(\mathbf{x})$. In the limit of large scales, this power spectrum should be related to the linear power spectrum of density perturbations as (see Croft et al., 1999, McDonald, 2003, McDonald et al., 2000)

$$P_F(k, \mu_k) = b_\delta^2 (1 + \beta \mu_k^2)^2 \, P_L(k) \, , \tag{2.1.1}$$

where μ_k is the cosine of the angle of the wavevector $\mathbf{k}$ in Fourier space relative to the line of sight, and P_L is the linear power spectrum of the mass density perturbations. This is the same form of the linear power spectrum derived by Kaiser (1987) for any class of observed objects with a bias factor b_δ, which relates the amplitude of observed fluctuations to the amplitude of the underlying mass fluctuations. But for the Lyα forest, the redshift distortion parameter β depends on a second bias factor that is related to the response of the mean value of F to a large-scale peculiar velocity gradient, and must be determined independently.

Therefore, the promise of massive spectroscopic surveys of Lyα absorption spectra is to help determine the shape of $P_L(k)$ over a wide range of scales and redshifts, and to use this to obtain crucial cosmological measurements, such as the angular and redshift scale of the Baryon Acoustic Oscillations, or the effect of neutrinos on the power spectrum (e.g., McDonald and Eisenstein, 2007). In addition, one can determine the values of b_δ and β at each redshift, which are in principle predictable with hydrodynamic simulations from the small-scale physics that determine the properties of the Lyα forest (McDonald, 2003). A first step in this direction was recently accomplished by Slosar et al. (2011) from the first analysis of the quasar absorption spectra in the BOSS survey.

Accurately measuring the power spectrum requires a careful evaluation and correction of any systematic errors that may be present in this measurement in the analysis of real data. The only way to reliably doing this is by generating several random realizations of the multiple Lyα absorption spectra in a survey, and introducing into them any possible systematic effects to see how they may impact the inferred power spectrum in the end. Some of the systematic effects that need to be considered are the following: errors in the modeling of the quasar continuum $C(\lambda)$, which is needed to evaluate the transmitted fraction from the observed flux, $f(\lambda) = C(\lambda)F(\lambda)$; variable spectral resolution and noise; flux calibration errors; the impact of the redshift evolution of the Lyα forest; the presence of damped Lyα , Lyman limit systems and metal absorption lines in the spectra; or variations in the intensity of the cosmic ionizing background. Modeling these systematic effects as accurately and reliably as possible requires our ability to generate mock surveys of Lyα absorption spectra in large numbers, for many different cases, and in a way that can be easily used. These mock surveys must include a large number of sources over large volumes (like the ongoing BOSS survey in SDSS-III; Eisenstein et al., 2011), and somehow include the small-scale fluctuations of the Lyα forest that are present in the observed spectra of sources that are point-like for practical purposes.

Generating these mock surveys directly from three-dimensional simulations, by selecting lines of sight from them, presents several difficult challenges. The first is that having a large enough volume to correctly simulate the power spectrum, at least up to scales as large as the BAO peak, implies that the resolution of the simulations cannot capture the smallest relevant scales for the Lyα forest. In addition, when using large three-dimensional simulations, the computer resources that are required may not allow obtaining many mocks that are independent, or changing the parameters of these mocks in an efficient and fast way to enable a large number of tests.

This paper presents a method to efficiently create these mock surveys of Lyα absorption spectra, taking advantage of the fact that the transmitted fraction F needs to be generated only on the discrete lines of sight to the survey sources. The method consists of generating one-dimensional fields for each line of sight and introducing correlations among them as if they had been drawn from a three-dimensional field. The capacity that is lost with this method is using hydrodynamic simulations that include the non-linear gravitational evolution of density fluctuations and other physical effects to simulate the field $F(\mathbf{x})$. However, if we care only about the large-scale power spectrum of this field and the errors to which it can be measured, it is in principle enough to ensure that the mocks have the same variance in the small-scale fluctuations to reproduce their effect on large scales. The way the mock surveys are generated is by using an input power spectrum of $F(\mathbf{x})$ in redshift space that includes a non-linear correction for small scales, and which is assumed to be calibrated from the results of cosmological simulations with enough resolution or directly from the observational results. The mocks can also include any one-point distribution of F that is desired and the redshift evolution of both the power spectrum and the distribution of F.

Hence, the philosophy of these mock surveys is that they are generated from an input model of the power spectrum and other quantities, and that they should be used for predicting the large-scale correlation measurements of the Lyα forest and the way they are affected by any systematic errors that can be introduced. However, the field $F(\mathbf{x})$ that is simulated is purely local and inferred from the linear overdensity, so it does not reproduce the 3-point or higher n-point correlations of the Lyα forest.

The method is presented in detail in §2, and an application to an example of a survey similar to BOSS is presented in §3. Another application of these mocks to simulate the effect of damped Lyα systems is discussed in Chapter 4. This method was already used for simulating the sample of spectra used in Slosar et al. (2011), and is being improved for application to the final BOSS survey.

A standard flat ΛCDM cosmology is used in this paper with the following parameters: $h = 0.72$, $\Omega_m = 0.281$, $\sigma_8 = 0.85$, $n_s = 0.963$, $\Omega_b = 0.0462$.

2.2 Method to Generate Mocks of Correlated Lyα Spectra

A Lyα forest spectrum is given by the fraction of transmitted flux, $F = \exp(-\tau)$, where τ is the optical depth, at each observed wavelength. We define the comoving coordinate in redshift space, x, related to the wavelength by $dx = c/H(z)(d\lambda/\lambda_\alpha)$, where $H(z)$ is the Hubble constant, the redshift is $1 + z = \lambda/\lambda_\alpha$ and $\lambda_\alpha = 1216$ is the Lyα resonance wavelength. The observed spectrum is the product of $F(x)$ times the continuum of the source, which is not independently observed and must be modeled. We shall not deal in this paper with the issue of modeling the continuum. Our mocks are realizations of the function $F(x)$ on multiple, correlated lines of sight.

In this paper we shall generally work with the variable

$$\delta_F(x) = \frac{F(x)}{\bar{F}} - 1 , \tag{2.2.1}$$

where $\bar{F}$ is the mean value of F at a given redshift. All the 2-point correlations appearing in this article are of this δ_F variable unless otherwise stated. This section describes the method to generate a set of mock Lyα spectra with any specified distribution function and power spectrum for the δ_F variable. The main idea for the case of a Gaussian field is explained in 2.2.1, which is then generalized to any desired distribution of δ_F (2.2.2). The inclusion of redshift evolution is discussed in 2.2.3.

2.2.1 Generation of a Gaussian Random Field

The most important requirement that our mock Lyα spectra must meet if they are to accurately predict any systematic and statistical errors in the measurements of large-scale correlations in δ_F is that they have a redshift space power spectrum of the flux that accurately matches the observed one. In this way, the intrinsic variance of the Lyα absorption at any scale can be reproduced, and the way it affects the sampling errors on all other scales is correctly taken into account. Our method to generate mock Lyα spectra can take as input any desired power spectrum $P_F(k_\parallel, k_\perp)$ in redshift space, where $k_\parallel$, $k_\perp$ are the components of the wave vector in Fourier space parallel and perpendicular to the direction of the line of sight.

Sampling the volume unevenly

The usual way to generate a Gaussian random field in realizations of cosmological perturbations is to generate first a set of independent Fourier modes in a three-dimensional cubic box with a specified power spectrum, and then doing the Fourier transform to obtain the real-space field. This method yields the value of the field at all the cells in the cubic volume at once.

However, to simulate the measurement of correlations up to the BAO scale in a survey of quasar spectra, we need to cover a volume with a size of at least several times the BAO scale, with a required resolution needed to capture the fluctuations in the low-density intergalactic medium of at least $\lambda_J/(2\pi) = \sqrt{3/2}c_s t$, or $\sim$ 100 comoving kpc (where λ_J is the Jeans length, c_s is the sound speed of the intergalactic gas, and t the age of the universe; see, e.g.,Peebles (1980)). The minimum dynamic range from the smallest to the largest scale is then $\sim 10^4$, or 10^{12} simulated points (and even larger if the entire volume of a survey like BOSS is to be generated), which results in a serious computational problem for being able to easily generate large numbers of mocks in a simple way.

Our method uses the fact that we are only interested in the values of the field along a number of infinitely thin lines of sight traced by the quasar light. Hence, we can generate a Gaussian field on these one-dimensional lines only, and introduce correlations among them directly in real space. A first, simple-minded way to achieve this might be to first generate an independent Gaussian variable at each pixel, g_i, and then combine them to generate the final field $\delta_{gj} = L_{ij} g_i$ which has the desired correlation C_{ij}:

$$C_{ij} = < \delta_{gi} \delta_{gj} > = < L_{ik} g_k L_{jl} g_l > = L_{ik} L_{jl} \delta_{kl} = L_{ik} L_{jk} \, . \tag{2.2.2}$$

A particularly efficient way to obtain the required matrix L for the transformation is the result of the Cholesky decomposition of the covariance matrix C, i.e., a lower triangular matrix L obeying $C = LL^T$. Numerically, there are several algebraic packages that perform the Cholesky decomposition very efficiently.

For a practical application, the number of pixels that are needed to model a typical observed spectrum and to include the power down to the smallest relevant scales is $N_p \sim 10^3$ for each line of sight. For a survey with N_q quasars, the total number of elements of the correlation matrix C that need to be computed is $(N_p \times N_q)^2$. Clearly, this method would break down for a relatively small number of quasars. Fortunately, there is a better way to do it.

Parallel lines of sight

Let us assume for the moment that the lines of sight in the survey are perfectly parallel. Let $\delta_g(x_\parallel, \mathbf{x}_\perp)$ be the correlated Gaussian variable we want to generate at the position $x_\parallel$ of the line of sight at coordinate $\mathbf{x}_\perp$. We can do the one-dimensional Fourier transform of δ_g on the direction of the line of sight only, to obtain $\tilde{\delta}_g(k_\parallel, \mathbf{x}_\perp)$. These one-dimensional Fourier modes have the following correlation:

$$\begin{aligned} \left\langle \tilde{\delta}_g \left(k_\parallel, \mathbf{x}_\perp\right) \tilde{\delta}_g \left(k'_\parallel, \mathbf{x}_\perp{}'\right) \right\rangle &= \frac{1}{2\pi} \int d\mathbf{k}_\perp \exp(i\mathbf{k}_\perp \mathbf{x}_\perp) \int d\mathbf{k}_\perp{}' \exp(i\mathbf{k}_\perp{}' \mathbf{x}_\perp{}') \\ &\times \delta^D \left(k_\parallel + k'_\parallel\right) \delta^D \left(\mathbf{k}_\perp + \mathbf{k}_\perp{}'\right) P\left(\mathbf{k}\right) \\ &= 2\pi \delta^D \left(k_\parallel + k'_\parallel\right) P_\times \left(k_\parallel, \left|\mathbf{x}_\perp - \mathbf{x}_\perp{}'\right|\right) \, , \end{aligned} \tag{2.2.3}$$

where the symbol δ^D stands for the Dirac delta function, $P(\mathbf{k})$ is the power spectrum of δ_g, and

$$P_\times \left(k_\parallel, r_\perp\right) = \frac{1}{2\pi} \int_{k_\parallel}^{\infty} k \, dk \, J_0 \left(k_\perp r_\perp\right) \, P\left(k_\parallel, k_\perp\right) \, . \tag{2.2.4}$$

The crucial property is that the one-dimensional modes $\tilde{\delta}_g$ on different lines of sight are independent except when $k_\parallel = k'_\parallel$. Therefore, the problem is now separated for each value of $k_\parallel$, and the Cholesky decomposition operation needs to be performed on N_p matrices of size $N_q \times N_q$ only.

Hence, the procedure to be followed in our method is as follows. We first choose a grid of values of $k_\parallel$ for the Fourier transforms on the line of sight. For each value of $k_\parallel$, we compute the correlation of the one-dimensional Fourier modes for every pair of lines of sight, using equations (2.2.3) and (2.2.4). Each one of these $N_q \times N_q$ matrices, $C_k = P_\times(k_\parallel, r_\perp)$, is then Cholesky-decomposed to obtain a matrix L_k. After generating a set of independent Gaussian variables for each quasar and each value of $k_\parallel$, g_{kq}, we compute the new set $\tilde{\delta}_g = L_k g$, and we then do the inverse one-dimensional Fourier transform of these to finally obtain the δ_g variables, with all the real space correlations that are implied by the input 3-d power spectrum $P(\mathbf{k})$.

In reality, the Lyα spectra need to be generated for quasars that are at different redshifts. We do this by first generating the spectra lines of sight of a long enough comoving length L, evaluating δ_g on bins of comoving width Δx. We set the center of the line of sight at a central redshift z_c (we use $z_c = 2.6$ in this paper), and every bin is then mapped into a redshift according to its comoving coordinate. We then use only the part of the spectrum of each quasar that is in the restframe wavelength range for Lyα forest analyses. We use 1041 Å$< \lambda_r <$ 1185 Å in this paper, the usual range to avoid Lyβ contamination and the proximity effect zone near the quasar. We also use $L = 4096\,h^{-1}$ Mpc, long enough to make any periodicity effects negligible, and $\Delta x = 0.5h^{-1}$ Mpc, slightly smaller than the typical pixel width in the BOSS spectrograph ($1 \simeq 0.7\,h^{-1}$ Mpc at the redshifts of interest).

2.2.2 Flux Distribution

The principal goal of the mocks of correlated Lyα forest spectra we want to generate is to simulate the observed spectra in a survey like BOSS that includes all of the statistical and systematic errors we may consider to obtain a correction for them when computing any statistical property. It is therefore important that the perturbation in the transmitted flux fraction, δ_F, in the mock spectra has the same distribution as the observed one, in order that the impact of continuum fitting and noise on the measured correlations and their errorbars are correctly simulated. Note that the value of the noise that is added in the mocks and the way that the continuum fitting is obtained will depend on a complex way on the values of δ_F. Here we generalize our method to generate a field δ_F with the desired probability distribution function $p_F(\delta_F)$ and any power spectrum

$P_F(\mathbf{k})$. Although the higher order n-point correlations of F will obviously still be different for the mocks and the real Lyα forest spectra, we expect this to have no impact on the computed errors of any statistical measurements on large scales.

This generalized method consists of generating first our field δ_g with a Gaussian distribution, $p_g(\delta_g) = \exp(-\delta_g^2/2)/\sqrt{2\pi}$, with a different power spectrum P_g such that, after transforming the field to the new variable $\delta_F(\delta_g)$, the desired probability distribution function $p_F(\delta_F)$ and power spectrum P_F are obtained. The required transformation $\delta_F(\delta_g)$ is obtained by integration of the equation

$$\frac{d\delta_F}{d\delta_g} = \frac{p_g(\delta_g)}{p_F(\delta_F)} \; . \tag{2.2.5}$$

Let us consider the correlation functions $\xi_F(r_{12})$ and $\xi_g(r_{12})$ of the field values at two points $\mathbf{x}_1$ and $\mathbf{x}_2$ separated by the distance r_{12}. We designate these field values as δ_{F1}, δ_{F2}, δ_{g1}, δ_{g2}. Since the field δ_g is strictly Gaussian, the correlation functions are related by

$$\begin{aligned}
\xi_F(r_{12}) &= \langle \delta_{F1}\, \delta_{F2} \rangle \qquad (2.2.6)\\
&= \int_{-1}^{1/\bar{F}-1} d\delta_{F1} \int_{-1}^{1/\bar{F}-1} d\delta_{F2}\, p_{2F}(\delta_{F1}, \delta_{F2})\, \delta_{F1}\delta_{F2}\\
&= \int_{-\infty}^{\infty} d\delta_{g1} \int_{-\infty}^{\infty} d\delta_{g2}\, p_{2g}(\delta_{g1}, \delta_{g2})\, \delta_{F1}\delta_{F2}\\
&= \int_{-\infty}^{\infty} d\delta_{g1} \int_{-\infty}^{\infty} d\delta_{g2} \frac{\exp\left[-\dfrac{\delta_{g1}^2 + \delta_{g2}^2 - 2\delta_{g1}\,\delta_{g2}\,\xi_g(r_{12})}{2(1-\xi_g^2(r_{12}))} \right]}{2\pi\sqrt{1-\xi_g^2(r_{12})}}\, \delta_F(\delta_{g1})\, \delta_F(\delta_{g2}) \; .
\end{aligned}$$

This relation between the two correlations ξ_F and ξ_g is actually a one-dimensional function that is totally independent of the separation r_{12} or any other variable: it depends only on the relation $\delta_F(\delta_g)$. We can therefore tabulate and invert the function $\xi_F(\xi_g)$.

The procedure to generate a random field δ_F is therefore the following: we start with an input model for the three-dimensional power spectrum P_F of the flux transmission, and compute the Fourier transform to obtain ξ_F. We then convert this to the correlation function ξ_g, and proceed to compute the correlations of one-dimensional power for the Gaussian field g in equation 2.2.4), which can be re-expressed as:

$$P_{g\times}(k_\parallel, r_\perp) = \int_{-\infty}^{\infty} e^{ik_\parallel r_\parallel} \xi_g(r_\parallel, r_\perp) \; . \tag{2.2.7}$$

We mention here that this procedure does not in general work for any distribution function $p_F(\delta_F)$, because sometimes the resulting power $P_{g\times}$ may be negative for some values of $k_\parallel$ and $r_\perp$. Fortunately, this does not occur for the input model chosen here, but it may well occur with other distributions (see Weinberg and Cole, 1992, for a discussion of the same problem in the context of non-gaussian initial conditions).

2.2.3 Redshift Evolution and Non-parallel Lines of Sight

The power spectrum of δ_F is a function of redshift. The main evolution is in the amplitude of the power spectrum, but a more general evolution in the shape is likely to be present, particularly on small scales. To introduce the redshift evolution in our model, we generate the field δ_F for several discrete values of the redshift, obtaining a set of realizations $\delta_{Fi}(x_\parallel, x_\perp)$, where the subindex i labels the redshift. Each of these realizations is generated with the same amplitudes and complex phases of the Fourier modes $\tilde{\delta}_g$, and varying only the amplitude of the power spectrum that is different due to the evolution with redshift.

The effect of the variation of the angular diameter distance and Hubble constant with redshift, and the fact that the lines of sight are not parallel, is included in the same way as the redshift evolution. The power spectrum can be expressed in terms of a fixed angular separation at the discrete values of the redshift at which the multiple fields δ_{Fi} are generated.

The final field δ_F is obtained by linear interpolation of the multiple fields as the redshift varies along the lines of sight, introducing in this way the gradual evolution in the power spectrum amplitude and the angular diameter distance with redshift.

In this paper, the redshift values at which the fields δ_{Fi} are generated are $z = 1.96$, 2.44, 2.91, and 3.39.

2.2.4 Input Model for Lyα Forest Mock Spectra

The distribution and power spectrum of the transmitted flux fraction can be determined from observations and can also be computed in theory from hydrodynamic cosmological simulations of the intergalactic medium. As observational progress is made, mocks of Lyα forest surveys can be adjusted to reproduce as accurately as possible the observational determinations of the distribution and power spectrum of δ_F, which guarantees an accurate modeling of the measurement errors for any quantities. Here, we use the parameterized fitting formula introduced by McDonald (2003) to fit the results of the power spectrum from several numerical simulations,

$$P_F(k, \mu_k) = b_\delta^2 (1 + \beta \mu_k^2)^2 P_L(k) D_F(k, \mu_k) \ , \tag{2.2.8}$$

where b_δ is the density bias parameter at $z = 2.25$, β is the redshift distortion parameter, $\mu_k = k_\parallel / k$, $P_L(k)$ is the linear matter power spectrum, and $D_F(k, \mu_k)$ is a non-linear term that approaches unity at small k. This form of P_F is the expected one at small k in linear theory, and provides a good fit to the observations reported in Slosar et al. (2011). Note that we do not generate a density and a velocity field, but we directly generate the Lyα forest absorption field instead, with the redshift distortions being directly introduced in the input power spectrum model of equation (2.2.8), with the free parameter β that measures the strength of the redshift distortion.

We use the parameters given in the central model of McDonald (2003), $b = -0.1315$ and $\beta = 1.58$ (the negative sign of b simply reflects the decrease of δ_F with gas density, and does not affect any equations in this paper because it always appears as b^2). Only the amplitude of the power spectrum is assumed to evolve with redshift, following a power-law:

$$P_F(k, \mu_k, z) = P_F(k, \mu_k, z = 2.25) \left(\frac{1+z}{1+2.25} \right)^{\alpha} . \tag{2.2.9}$$

We use the value $\alpha = 3.8$ in this paper, as suggested by the evolution of the one-dimensional P(k) measured in McDonald et al. (2006).

For the probability distribution, we use a log-normal model for the optical depth τ,

$$F = e^{-\tau} = \exp\left(-a e^{\gamma g}\right) , \tag{2.2.10}$$

where g is a Gaussian variable of unit dispersion, and a and γ are two free parameters determining the mean transmission $\bar{F}$ and its variance.In the future, a new distribution for F that more accurately matches the observed one should be used for the mocks, but the log-normal approximation suffices for the purpose of this paper of demonstrating the applications of Lyα forest mocks.

We assume a mean transmitted fraction that approximately matches the observations, (McDonald et al., 2006):

$$\ln \bar{F}(z) = \ln(0.8) \left(\frac{1+z}{3.25} \right)^{3.2} . \tag{2.2.11}$$

The values of a and γ at each redshift can be derived by requiring the mean value of F to match equation (2.2.11), and the dispersion to reproduce the value implied by the power spectrum P_F. The result for the parameters at the four redshifts we use are the following: $a = 0.065$ and $\gamma = 1.70$ at $z = 1.96$; $a = 0.141$ and $\gamma = 1.53$ at $z = 2.44$; $a = 0.275$ and $\gamma = 1.38$ at $z = 2.91$; and $a = 0.487$ and $\gamma = 1.24$ at $z = 3.39$.

2.3 Results

This section presents the results for the characteristic errors in the measurement of the correlation function, in an example of a simulated Lyα forest survey with similar characteristics as BOSS.

2.3.1 Model for the Quasar Survey

The first step to generate a mock Lyα forest survey is to generate the quasar sample. We randomly distribute quasars (with no clustering) over a circular area $A = 300\,\mathrm{deg}^2$ and the redshift range $2.15 < z < 3.5$, following the quasar luminosity function measured in Jiang et al. (2006) up to a limiting magnitude of $g = 22$. We select only 75 % (independently of g magnitude and redshift) of the quasars in order to have a quasar number density closer to the one obtained in the BOSS survey ($\sim 15 - 17\,\mathrm{deg}^{-2}$). The total number of quasars in the sample is $N_q \simeq 5000$. The code we use to generate the absorption fields with the method described in Section 2 was able to generate all the absorption spectra in one survey mock with a node with 8 CPU in a few hours.

The redshift distribution of the sources in a real survey usually differs substantially from that inferred from the model luminosity function, mainly because the target selection efficiency has a strong dependence on redshift. In particular, in the optical color selection used by SDSS, quasars at $z \sim 2.7$ overlap the stellar locus and are confused with stars, making them harder to select. There is also a change in efficiency as a function of the foreground stellar density and dust absorption. We do not include these effects here. If anything, these effects should reduce the errors of measuring the Lyα correlation because they should cause an increased overlap of the Lyα spectra redshift range and an increased number of quasar pairs at small separations, for fixed mean quasar density.

After having constructed the spectra of the transmitted fraction F as described in the previous section, we generate a realistic observed quasar spectrum that includes a spectral resolution and noise approximately matching those in the BOSS survey, following these steps:

- A new set of pixels for a mock of the physical spectrum in units of flux is constructed, covering the whole, fixed wavelength range 3600 Å$< \lambda <$ 9000 Å, with pixels of constant wavelength width $\Delta\lambda = 1$ Å. The width of these pixels in comoving separation is therefore changing along the spectrum.

- For each quasar, we compute the mean value of the pixel width in comoving separation, $\langle\Delta\chi\rangle$, over the region that is used for measuring the Lyα forest correlation function, 1041 Å$< \lambda <$ 1885 Åin the rest frame. We then convolve our spectrum of δ_F in the original pixels of constant comoving length with the Point Spread Function that results from the convolution of a Gaussian spectral resolution and the pixel width in the final wavelength bins:

$$\delta_F(x) = \frac{1}{2\pi}\int dk\, e^{ikx}\tilde{\delta}_F(k)\ \exp\left[-\frac{k^2\langle\Delta\chi\rangle^2}{2}\right]\ \left[\frac{\sin(k\langle\Delta\chi\rangle/2)}{k\langle\Delta\chi\rangle/2}\right]^2 . \qquad (2.3.1)$$

The value of $\Delta\chi$ depends on the quasar redshift, with its typical value being in the range $0.6 - 0.8\,h^{-1}$ Mpc. We note that the wavelength dependence of the spectral resolution and the pixel width in the BOSS spectrograph are actually quite complex, and they should be carefully treated if one is interested in small-scale correlations.

- Each pixel in the spectrum with constant wavelength bins is assigned the value of F in the nearest bin of the spectrum with pixels of constant comoving width. We set $F = 1$ for wavelengths outside the Lyα forest range.

- We multiply the spectrum of F by the continuum for each quasar, using the mean rest-frame spectra obtained in Suzuki et al. (2005). A spectrum of physical flux, $f(\lambda)$, is obtained after normalizing to match the g magnitude of the quasar.

- The expected noise variance for the case of the BOSS spectrograph with an exposure time of 1 hour is computed at each pixel using the expression

$$\sigma_N^2(\lambda) = A + B(\lambda)\ [f(\lambda) + s(\lambda)]\ \Delta\lambda\,, \qquad (2.3.2)$$

  where $s(\lambda)$ is a typical sky flux in BOSS, A is the read-out noise and $B(\lambda)$ is related to the BOSS throughput. These functions have been kindly provided by David Schlegel.

- We add a Gaussian random variable with variance σ_N^2 to the flux $f(\lambda)$ at each pixel, and divide the resulting flux by the continuum to obtain a new spectrum of transmitted fraction F (which is no longer restricted to the range $0 < F < 1$ because of the noise that has been added).

The detailed properties of the noise in the real survey are more complicated, but this simple procedure allows us to approximately study the effect of noise on the correlation function measurement.

An example of mock spectra with continuum and noise added is shown in Figure 4.3.

We have generated 50 realizations of this mock survey to obtain the results that are presented next.

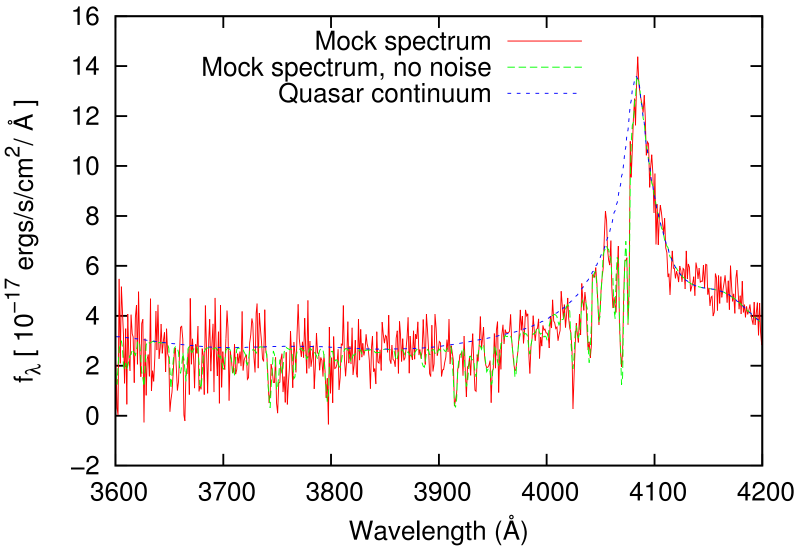

Figure 2.1: Mock quasar spectrum (red), without noise (red) and without Lyα absorption (blue).

2.3.2 Measurement of the Correlation Function

We estimate the value of the correlation function as the weighted average of the product of the δ_F variable in all pixel pairs that have a redshift space separation r, angle cosine μ and mean redshift z, which are within a certain bin of width Δr, $\Delta \mu$ and Δz, which we designate as A:

$$\hat{\xi}_A = \frac{\sum_{i,j \in A} w_i w_j \, \delta_{Fi} \delta_{Fj}}{\sum_{i,j \in A} w_i w_j} \, . \tag{2.3.3}$$

To calculate the correlation function from a mock survey, we initially use a very large number of bins, with a total of 150 bins in r up to $r = 150\, h^{-1}\,\mathrm{Mpc}$, 20 bins in μ and 20 bins in z (all of them linearly spaced). These bins are thin enough for the final results to have converged to the correct value in the limit of small bins. The weights are set equal to the total inverse variance in each pixel, including the intrinsic Lyα forest fluctuations, $\sigma_F^2(z) = \langle \delta_F^2 \rangle$, and the variance caused by the instrumental noise, $\sigma_N^2(\lambda) / \left[\bar{F(z)} C(\lambda) \right]^2$,

$$w_i = \sigma_i^{-2} = \left[\sigma_F^2(z_i) + \frac{\sigma_N^2(\lambda_i)}{\left(\bar{F(z_i)} \; C(\lambda_i) \right)^2} \right]^{-1} \, . \tag{2.3.4}$$

This ignores the effect of the intrinsic variance correlation in neighboring pixels for the purpose of computing the optimal weights assigned to each pixel, an approximation that was also used in Slosar et al. (2011).

The noise variance in equation 2.3.2 applies to the flux variable $f(\lambda) = \bar{F}[1 + \delta_F(\lambda)] \, C(\lambda)$. The corresponding contribution to the variance of δ_F in equation 2.3.4 is obtained by dividing by $\left[\bar{F(z)} C(\lambda) \right]^2$.

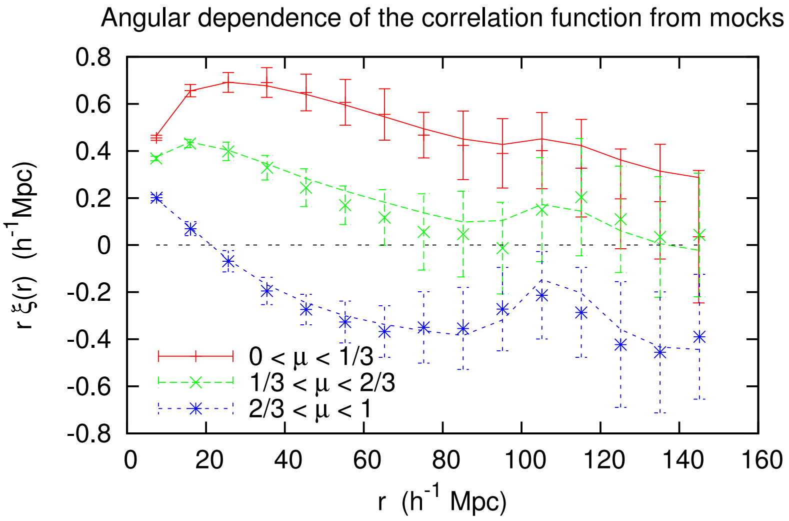

Figure 2.2: Correlation function obtained from the average of 50 mocks of our survey model (points), as a function of r, in three diferents bins in $\mu = cos(\theta)$. Errorbars show the dispersion of a single mock. The curves are the expected correlation function from the input model after averaging in the same bins.

The correlation function is then averaged over all redshifts and compressed into broader bins of 5 Mpc$/h$ in r, and three bins in μ. The results are plotted in Figure 2.2 as the points with errorbars. The size of the errorbars show the dispersion in one realization, as computed from the 50 independent realizations of the survey that we generate, while the points indicate the mean of all 50 realizations. For example, the BAO feature is not significantly detectable from one survey like the one we simulate, as indicated by the size of the errorbars, but is detected at high significance from the average of all 50 realizations.

Careful inspection of this figure shows that the agreement between the input theory and the mean measurement from the mocks is not perfect. For example, the points in the range 50 to 100 Mpc$/h$ in the middle bin of μ are systematically low compared to the theory by three to four times the dispersion in the mean of all realizations (the dispersion in the mean is a factor $\sqrt{50-1} = 7$ smaller than the dispersion in a single mock, shown as the errorbars). Unfortunately, we have not been able to eliminate this discrepancy, which must arise from a numerical inaccuracy in the numerous operations that are needed to go from the input power spectrum of δ_F, to the correlated modes to generate δ_g and then transform it back to δ_F, to the evaluation of the correlation function from the generated mocks by counting of pixel pairs. This small discrepancy would become important (comparable to the statistical errors from a simulated survey) when analyzing datasets larger than the first-year BOSS dataset analyzed in Slosar et al.

(2011). We expect to solve this residual numerical discrepancy in a future revision of the numerical code to generate mock surveys of Lyα spectra.

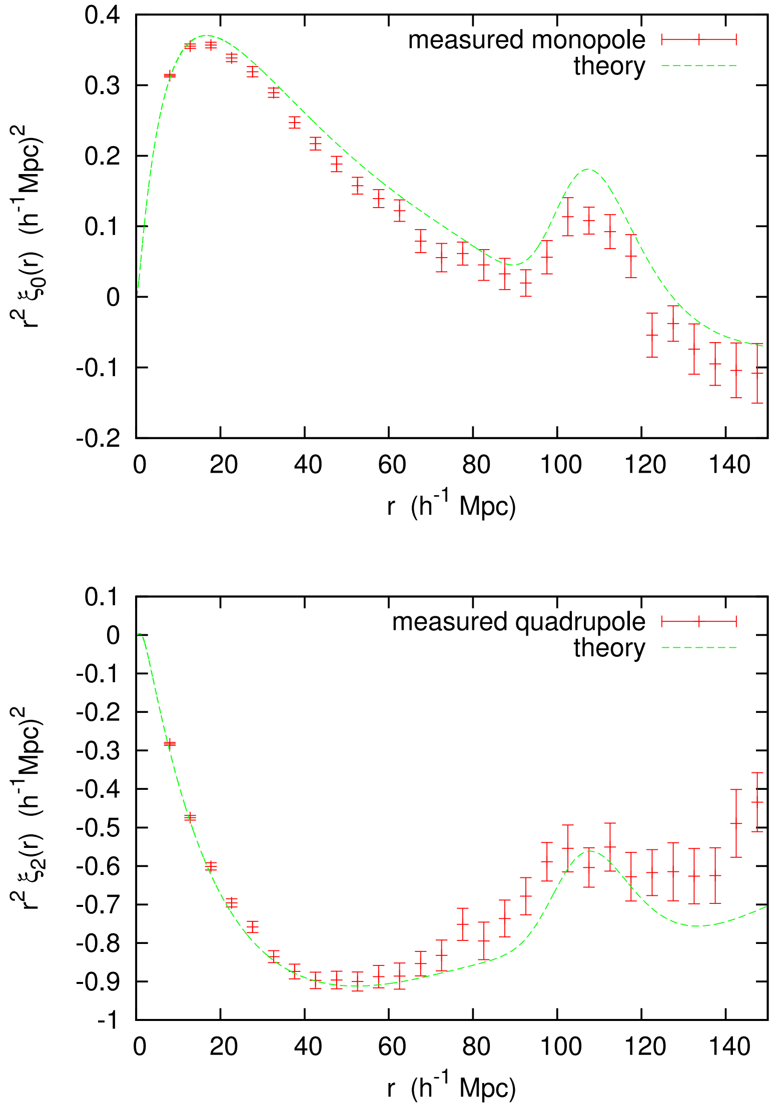

Figure 2.3: Monopole and quadrupole of the correlation function obtained from the mocks (points), compared to the input model (solid curve). The errorbars here have been rescaled to mimic a survey of the same size as BOSS.

The angular dependence of the correlation function can also be measured using the multipoles. The monopole and quadrupole of the average of the 50 mocks are shown in Figure 2.3 as the points. The green solid curve shows the input model for the correlation function, computed as before from the Fourier transform of the power spectrum model in equation (2.2.8), and averaging over the same bins in r, μ and z before computing the monopole and quadrupole.

The errorbars in this case have been rescaled to mimic the expected errors in the full BOSS survey. The area of the BOSS survey will be approximately 30 times that of our model survey described in 3.1, with approximately the same quasar density, so we simply reduce the errorbars of one of our mocks by a factor $\sqrt{30}$. This neglects the edge effects of the survey (the fact that fewer pairs of quasars are found for quasars near the edge of the survey area), which are small. According to this prediction, the amplitude of the BAO peak should be detectable at the $\sim 5-\sigma$ level in $5h^{-1}$ Mpc bins in the correlation function if all the data obtained in BOSS is as expected.

2.3.3 Variations in the Survey Strategy

An application of the mock Lyα forest surveys is to calculate the precision achieved in the measurement of the correlation function on large scales as a function of any survey properties in order to optimize the design of the survey. This study may often be done using a Fisher matrix approach without the need to generate survey realizations, but using the mocks presented here allows one to include any possible systematic effects in a more complete way.

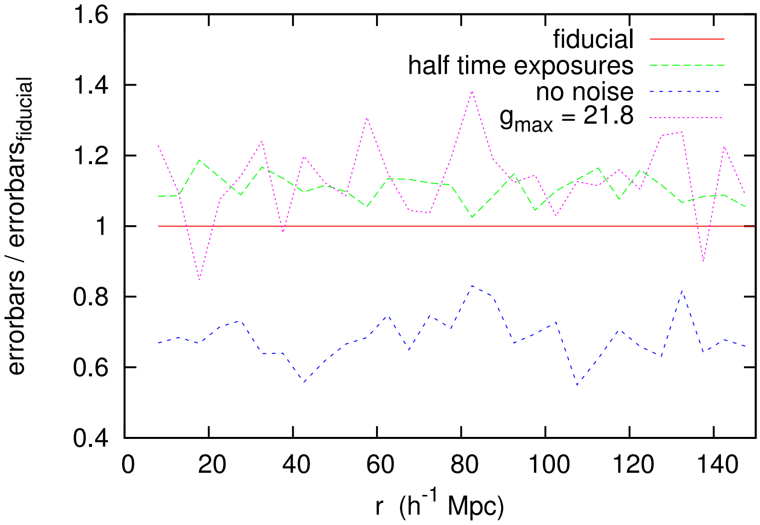

Figure 2.4: Fractional change in the errorbars of the monopole of the correlation function, for each radial bin, with respect to the fiducial survey, when varying survey parameters. The dashed green line assumes that all exposure times are divided by two, and the dotted pink line shows the result of eliminating the faintest 16% of the quasars with $21.8 < g < 22$. The dotted blue line is for the case with no observational noise.

Here we study the change in the errorbars of the monopole of the correlation function when we vary either the exposure time or the number of observed quasars within a fixed area. We note that the variation of these errorbars with the area of the survey, if we keep the quasar density fixed, is basically proportional to the inverse square root of the area, apart from the presence of edge effects, which are already small at the BAO scale for our fiducial survey with an area of 300 deg^2.

Figure 2.4 shows that the fiducial survey has errors that are reduced by $\sim$ 30 % if the observational noise (both photon and read-out noise in the detectors) were entirely eliminated. In other words, the errors arising from observational noise and from the intrinsic sampling variance in the Lyα forest are comparable in our fiducial survey. The best strategy to reduce the sampling variance is to aim for the largest possible survey area. Increasing the source density is more difficult because one has to search for fainter quasars, which are harder to identify and have larger observational noise for a fixed exposure. The curves in Figure 5 show that reducing the exposure time by a factor of 2 degrades the errorbars by the same amount (10 to 15%) as eliminating the faintest 16% of the quasars, in the magnitude range $21.8 < g < 22$. Therefore, this shows that maximizing the number of quasars that are observed is the best survey strategy, even near the magnitude limit of the BOSS spectroscopic quasar survey (see Ross et al., 2011), and even if this is done at the cost of some reduction in the exposure time.

McDonald and Eisenstein (2007) used a simple Fisher matrix approach to study the best survey strategy to measure the angular diameter distance $D_A(z)$ and the Hubble parameter $H(z)$ from the BAO peak in the correlation function. In their Figure 1, these authors show that when the survey limiting magnitude is reduced from $g = 22$ to $g = 21.8$, the fractional error on the angular distance $D_A(z)$ increases by $\sim$ 20% and the Hubble parameter $H(z)$ increases by $\sim$ 10%, in agreement with the $10 - 15$% increase of the errorbars that we find (the S/N used for their figure is higher than in our mocks, so their improvement for a fainter limiting magnitude should be slightly higher than ours). In their Figure 5, McDonald and Eisenstein (2007) show that the fractional error on both scales increases by $\sim$ 10% if the $(S/N)^2$ is reduced by a factor of 2 (equivalent to reducing the exposure time by a factor 2), also in agreement with the $10 - 15$% increment of the errorbars found in the analysis of our mocks.

Our method is highly flexible to allow for a rapid computation of the best strategy for survey optimization, including any systematic effects that one may consider and include in the mocks in a realistic way.

2.4 Conclusions

The method described here is able to create mock correlated spectra of Lyα forest surveys mimicking the observed properties. Two free functions can be introduced as input to the mocks, fixing the one-point distribution and two-point correlation function of the field δ_F, which can be made to evolve with redshift. The higher order n-point functions that are not reproduced are assumed to not affect the measurements of 2-point statistics on the large scales of interest.

This paper presents only a simple example of the application of these mocks to a survey with similar characteristics as BOSS. The technique has already been used in the first analysis of BOSS data in Slosar et al. (2011). In the future, we plan to improve our methodology to use it on a number of sources as large as the entire BOSS survey, and to include all the observational effects in increasing detail. One of the main applications of these mocks is to accurately model the effect of high column density systems and metal-line absorption systems on the measurement of the Lyα forest correlation, which will be described in Chapter 4.

CHAPTER 3

The Lyman-α Forest in Three Dimensions: Measurements of Large-scale Flux Correlations from BOSS 1st-year Data

3.1 Introduction

My thesis work has been taken place in close connection to my participation in the BOSS survey of the SDSS III collaboration.

In May 2011 the Lyα forest working group of the Baryon Oscillations Spectroscopic Survey (BOSS) published a paper by Slosar et al. (2011) on the large scale correlation of Lyα absorption from the first year of BOSS data.

Several studies had previously detected correlations between close pairs of quasars (Bechtold et al., 1994, Dinshaw et al., 1994, 1995, Smette et al., 1992). However, Slosar et al. (2011) is the first one in which the correlation is measured on cosmologically large scales at which the fluctuations are close to linear. These first results from the BOSS survey support our expectation of detecting the Baryon Acoustic Oscillations in the correlations of the Lyα forest, which is the main goal of the survey.

This study, lead by Dr. Anze Slosar from the Brookhaven National Laboratory (US), is the work of the entire Lyα working group in BOSS. I therefore do not include it as a chapter of my thesis, but is included in appendix B. Here I write my own summary of this paper of the SDSS-III collaboration, and I describe my close participation in this work, in particular with Dr. Slosar, Dr. McDonald and Dr. Miralda-Escudé. Finally I focus on the role played by the mock catalogs that I have developed on the results of this paper.

3.2 BOSS Collaboration

The Baryon Oscillations Spectroscopic Survey (BOSS) is one of the four surveys of the third phase of the Sloan Digital Sky Survey (SDSS) (Eisenstein et al., 2011). The main goal of the survey is to study the nature of Dark Energy (DE) and the geometry of the universe by measuring the Baryon Acoustic Oscillations (BAO) signature imprinted in the clustering of galaxies and in the correlations of the Lyα forest at higher redshifts.

SDSS (Fukugita et al., 1996, York et al., 2000) has been one of the most influential surveys in the history of astronomy. Observations started in 2000, with a full-time dedicated 2.5 meter telescope located in the Apache Point Observatory, New Mexico (US).

During the first phase of operations (2000-2005), the telescope surveyed more than 8000 square degrees in five photometric bands (u-g-r-i-z). Spectra were obtained for thousands of galaxies and quasars, providing the largest catalog of galaxy redshifts at the time.

During the second phase (2005-2008) the SDSS primary goals were completed. In addition, two new projects were included in the survey: SEGUE (Sloan Extension for Galactic Understanding and Exploration) probed the dynamics and history of our galaxy, and the Sloan Supernova Survey repeatedly observed a stripe of 300 square degrees to study variable objects, discovering nearly 500 confirmed Type Ia supernovae.

The third phase of the survey consists of BOSS and 3 other projects: SEGUE-2 continues the exploration of the Milky Way structure that was started with SEGUE. APOGEE (APO Galactic Evolution Experiment) is obtaining infrared spectra of thousands of red giant stars in the Milky Way to study the dynamics and chemical evolution of the galaxy. Finally MARVELS (Multi-object APO Radial Velocity Experiment Large-area Survey) searchs for giant gas planets in 11 000 nearby bright stars.

In order to measure the BAO imprint on the large scale structure, BOSS plans to obtain spectra of 1.5 million galaxies to redshift $z = 0.7$ and of 160 000 high redshift quasars ($z > 2.1$). This is done by using a new multi-object spectrograph that observes 1000 objects simultaneously, over the wavelength range 360 nm $< \lambda <$ 1000 nm, and with a mean resolution of $R \sim 2000$.

BOSS will measure the correlation function of galaxies and Lyα absorption, which should contain the BAO bump as an imprint from the recombination epoch.

By using the BAO scale as a standard ruler, BOSS will measure both the angular distance $d_A(z)$ and the Hubble parameter $H(z)$ with few percent precisioni over the redshift range $0.3 < z < 0.7$ using galaxies, and $2.0 < z < 3.5$ using the Lyα forest. These measurements will shed light on our understanding of both the nature of the accelerated expansion of the universe and its geometry.

The survey that started in the Fall of 2009 is designed to be completed by the Spring of 2014. The first public data release with BOSS spectra is planned for July 2012 and will contain all the spectra obtained over the first 2 years of observations. Different working groups within the collaboration have been working extensively with the data, and two first publications already appeared in early 2011: White et al. (2011) measured the clustering of 44 000 galaxies obtained during the first half year of observations, and Slosar et al. (2011) measured the correlation function of the Lyα forest in 15 000 quasars using the first year data.

Even though the Lyα forest had previously been used as a cosmological tool, most of the previous studies had employed the 1D power spectrum from individual lines of sight, or the correlation between close pairs of quasars. In this study, we detect the correlation across widely separated lines of sight, confirming the cosmological nature of the Lyα forest and supporting the forecasts of BAO detection using the Lyα forest.

Since the first publication of the Lyα forest clustering, the activity of our working group is veigorously engaged in the continuing scientific effort to analyze the data as it is obtained. The working group includes more than 40 members from several different countries (Spain, France, US, Italy and Germany).

3.3 Review of Slosar et al. 2011

This section describes the first detection of correlations in the Lyα absorption across widely separated lines of sight. Using the first year data from the BOSS survey containing roughly 15 000 quasars, we measure the correlation function up to $r = 100\,h^{-1}$ Mpc. We show that a linear bias model accurately describes the correlations on large scales $r > 10\,h^{-1}$ Mpc.

We now describe the data sample used in subsection 3.3.1, the methods used in the data analysis in 3.3.2 and the main results in 3.3.3. We discuss the results and present some conclusions in subsection 3.3.4.

3.3.1 Data Sample

BOSS targets a mean number of $\sim$ 40 objects as quasar candidates per square degree, following a complex target selection method described by Ross et al. (2011). Half of these candidates are selected from the SDSS photometry with a method that is uniform in the whole survey (CORE sample). The other half exploits a lot of additional information (mainly variability and UV or infrared photometry) in order to maximize the number of quasars that are found (BONUS sample), without attempting to make the sample uniform. The uniform CORE sample can be used for quasar clustering studies, which require the selection function to be well known and uniform while the BONUS sample is used only for the Lyα forest studies.

Roughly 15 to 20 out of the 40 targets per square degree are expected to be high-z quasars ($z > 2.1$), while the are mostly stars and low - z quasars. These are not useful for Lyα forest studies because the forest lies outside the range of the spectrograph.

During the first year of observations, the target selection method was not yet optimal, so only $\sim$ 15000 out of the $\sim$ 50000 targets were high - z quasars. Since then, the success rate of the BOSS survey has improved mostly because of the use of additional information from UV and infrared data, and because of the better weather enjoyed at Apache Point in the 2010-11 season.

All targets were visually inspected by the French Participation Group (FPG). Warning flags for Broad Absorption Lines (BAL) quasars were assigned to $\sim$ 1300 objects. The FPG inspection also detected a similar number of quasars with Damped Lyman α systems. In this study we discard all quasars flagged as BAL, and we use only those flagged as DLA when explicitly stated.

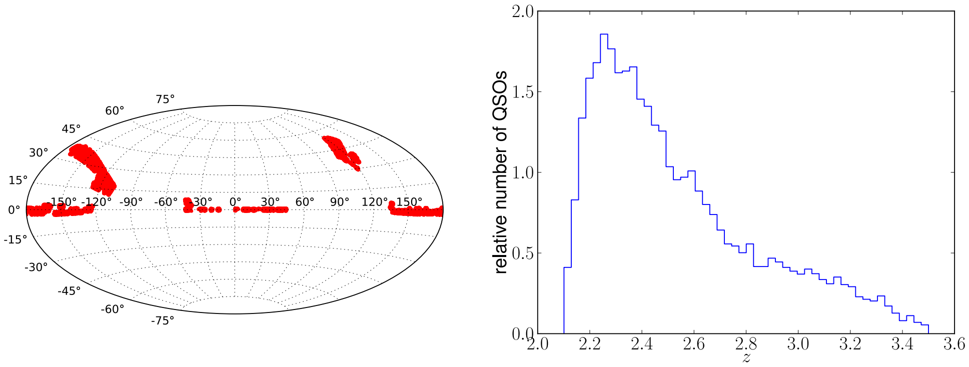

Figure 3.1: Distribution of observed quasars in sky position (left, in Aitoff projection) and in redshift (right).

The sky area inspected during this first year is shown in figure 3.1, together with the redshift distribution of the quasars used in this study.

3.3.2 Data Analysis

The raw BOSS data from each plate is a set of spectra recorded in a single CCD image, which contains one spectrum for each of the fibers. The standard SDSS pipelines spectro2d and spectro1d reduce the data into a single spectrum for each object, after standard operations of sky substraction, flux calibration and addition of several separate exposures.

The first step to extract the Lyα signal is to fit a continuum to the spectra, while measuring the mean transmitted flux as a function of redshift. In this study we assume that all quasar continua have the same shape, shifted horizontally as a function of quasar redshift z_i and vertically as a function of quasar magnitude, with an allowed extra tilt to correct for spectophotometric errors. Our model for the measured flux for a pixel with wavelength λ of quasar i is:

$$f(\lambda, i) = a_i \left[\lambda_r/(1185\,)\right]^{b_i} C(\lambda_r)\bar{F}(\lambda) \left[1 + \delta_F(\lambda, i)\right] \ , \tag{3.3.1}$$

where $\lambda_r = \lambda/(1+z_i)$ is the rest-frame wavelength, $\bar{F}(\lambda)$ is the mean transmitted flux fraction at redshift $z = \lambda/\lambda_\alpha$ and $C(\lambda_r)$ is the mean quasar continuum, which is multiplied by a power law $a_i[\lambda/(1185\,)]^{b_i}$ determined for each individual quasar. The variable used in the correlation analysis is $\delta_F(\lambda, i)$.

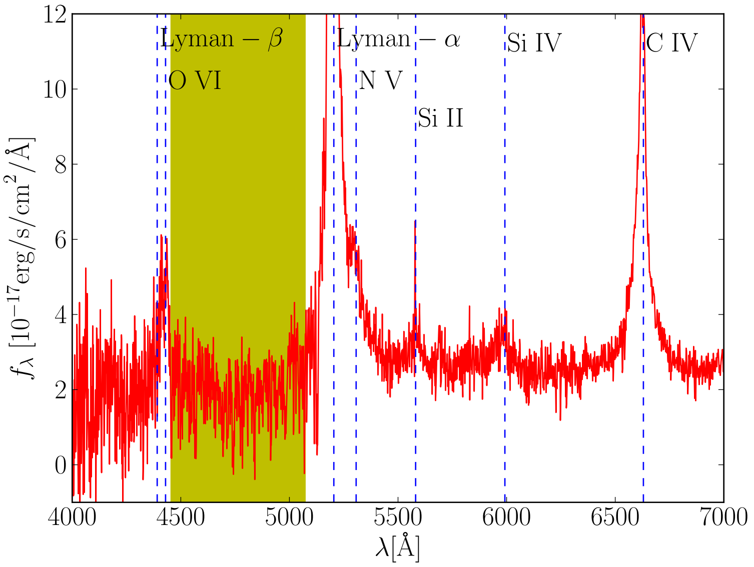

Figure 3.2: A quasar spectrum at redshift $z = 3.276$ from the first year of BOSS data. The Lyα forest is the shaded region between the Lyα and the Lyman-β emission lines. Other strong emission lines are also indicated.

For this analysis we define the Lyα forest as the rest-frame wavelength range 1041 Å$< \lambda_r < 1185$ Å, as plotted in figure 3.2. In order to decrease the variance in our measurements we force the mean value of δ_F to be 0 in each spectra, where the mean is computed only in the Lyα forest region and the pixels are weighted with the total variance in each pixel. This erases all modes along the line of sight with wavelength larger than the size of the forest. We model this effect when comparing the measured data with the theory.

We estimated the value of the correlation function in 12 radial bins up to $r = 100\,h^{-1}$ Mpc, 10 angular bins in $\mu = cos(\theta)$ and 3 redshift bins ($z < 2.2$, $2.2 < z < 2.4$ and $z > 2.4$, where the redshift refers to the absorbing gas, not the quasar). For each bin we compute the correlation function as a weighted average over pixel pairs,

$$\bar{\xi}_F(r,\mu) = \frac{\sum_{\text{pairs } i,j} w_i w_j\, \delta_{Fi}\delta_{Fj}}{\sum_{\text{pairs } i,j} w_i w_j}\ , \tag{3.3.2}$$

where the weights w_i are the total inverse variance,

$$w_i = \sigma_i^{-2} = \left[\sigma_{Fi}^2 + \sigma_{Ni}^2\right]^{-1}\ , \tag{3.3.3}$$

including both the intrinsic variance of the forest σ_{Fi}^2 and the variance due to observational noise σ_{Ni}^2. To avoid introducing spurious correlations that may result from a bad continuum fitting, we do not use pixel pairs from the same quasar.

Data	bias	β	$b(1+\beta)$	α
$r > 20\,h^{-1}$ Mpc	0.197 ± 0.021	$0.71\pm^{0.21\ 0.49\ 0.87}_{0.16\ 0.27\ 0.39}$	0.336 ± 0.012	1.59 ± 1.55
$r > 10\,h^{-1}$ Mpc	0.175 ± 0.012	$0.90\pm^{0.15\ 0.33\ 0.56}_{0.13\ 0.23\ 0.33}$	0.333 ± 0.008	2.09 ± 0.94
With LLS/DLA, $r > 20\,h^{-1}$ Mpc	0.217 ± 0.025	$0.55\pm^{0.19\ 0.48\ 0.97}_{0.14\ 0.25\ 0.35}$	0.337 ± 0.014	2.99 ± 1.74
With LLS/DLA, $r > 10\,h^{-1}$ Mpc	0.180 ± 0.013	$0.87\pm^{0.16\ 0.35\ 0.56}_{0.13\ 0.25\ 0.35}$	0.337 ± 0.009	3.11 ± 0.93

Table 3.1: This table shows the results of the parameter fittings. All error bars are $1-\sigma$ error bars, except for the β parameter in which case we give 1,2 and $3-\sigma$ confidence limits.

Fitting a parametrized theoretical model to the data requires knowledge of the covariance matrix of our measurements of the correlation function ξ_F in different bins. This is a non-trivial problem owing to the large number of bins that need to be used and the difficulty to obtain a reliable covariance matrix from simulations. We solved this problem by directly computing an estimate of the covariance from the very same data, exploring with a Monte Carlo procedure the pairs of pixel pairs that have a larger contribution. We tested the procedure with the mocks described in the previous chapter.

We finally fitted the data with a simple linear bias model, where the correlation function is simply the Fourier transform of the redshift distorted linear power spectrum,

$$P_F(k,\mu,z) = b^2(1+\beta\mu^2)\; P_L(k, z=2.25) \left(\frac{1+z}{1+2.25}\right)^{\alpha} , \tag{3.3.4}$$

where we have modeled the redshift evolution as a simple power law. The whole data analysis code has also been extensively tested with the mock catalogs described in the previous chapter.

3.3.3 Results with the Observed Data

We detect the correlation up to a distance $r = 60\,h^{-1}$ Mpc with a significance of $3-\sigma$, and up to $r = 70\,h^{-1}$ Mpc at $2-\sigma$.

The main results are captured in figure 3.3, where the observed monopole is plotted (averaged over redshift), together with the observed correlation in the $r_\perp - r_\parallel$ plane, compared to the best fit theory. The best-fit parameters are shown in table 3.1. If we compare these values with the expected parameters from the theoretical work of McDonald (2003), the measured value of b is somewhat large whereas the value of β is somewhat low. Work by Other authors has actually predicted values closer to the observations, but the measured values can be severely affected by the presence of high column density systems and metals, as described in chapter 4.

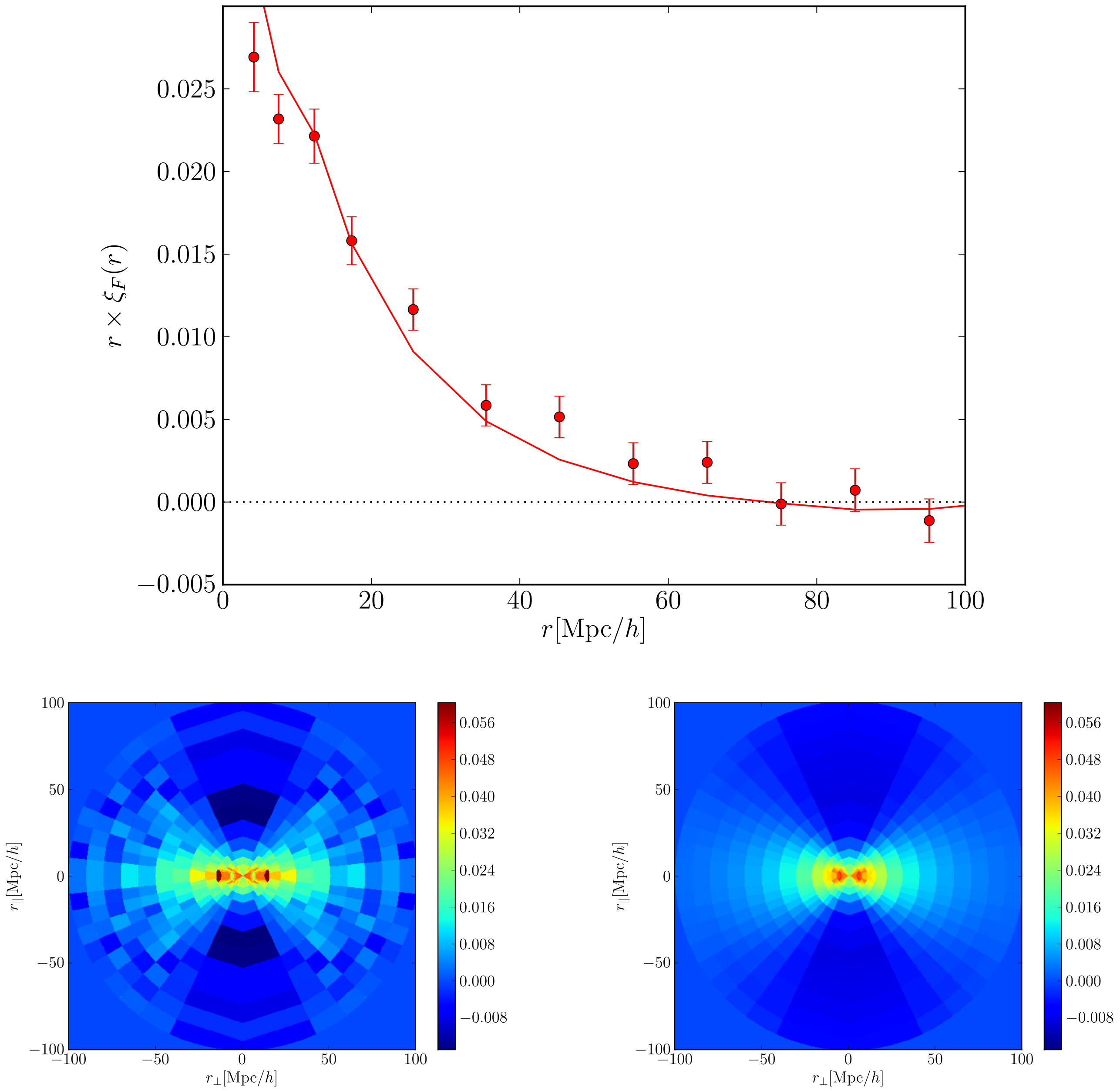

Figure 3.3: Primary measurement results in visual form. Top plot shows the monopole of the correlation function, together with a best-fit two-parameter (b, β) linear model. The bottom two plots are redshift averaged data plotted in the plane $r_\perp - r_\parallel$, with each pixel plotted with the value corresponding to the nearest neighbor. The left panel corresponds to data, and the right panel to the corresponding best-fit theory.

We also investigated the stability of the previous results by splitting the data in different samples: quasar magnitude, rest-frame wavelength and redshift. None of the tests resulted in a detection of any significant systematic effect.

3.3.4 Conclusions and Prospects

The analysis of the first Lyα forest BOSS data is the first probe of the cosmological origin of the Lyα forest on large scales. The fact that we are able to fit the whole data set with 330 bins using a linear bias model with only 3 parameters (b, β, α) supports the prospects of using the Lyα forest clustering to measure the BAO scale at high redshift.

There are several possible improvements to the methods that should be investigated before the next study is undertaken. One of these improvements is the impact of possible systematics such as the presence of high column density systems and metals.

3.4 Personal Contribution

I joined the BOSS collaboration in January 2009, when I was visiting Patrick McDonald at the Canadian Institute for Theoretical Astrophysics (CITA). I decided to focus my participation on the construction of mocks of the whole BOSS Lyα forest survey.

In the first meeting of the collaboration (Ohio, March 2009) I presented the first results in our attempt to generate realistic mock catalogs of the BOSS quasar spectra, in collaboration with Patrick McDonald and my PhD advisor Jordi Miralda.

As the code to generate mock catalogues evolved, different versions of the mock data were released to the collaboration, and the feedback from the people using them helped to improve the code while my cross-checks of the data analysis helped to develop the code that was being build to analyze the data.

In October 2010, after the SDSS-III meeting held in Paris in late September, a core team was created (Anze Slosar, Patrick McDonald, Jordi Miralda, Jim Rich, Jean-Marc LeGoff, Matt Pieri, Nicolas Busca and myself) to carry on a first study of the Lyα forest clustering with the first year of data, that ended with the publication of Slosar et al. (2011) on May 2011.

3.4.1 Role of the Mock Catalogs in the Publication

This study represents the first attempt to measure the correlation function of the Lyα forest across multiples lines of sight. New techniques have been developed in order to analyze the data, estimate the uncertainties and test different possible systematics.

Moreover, this has also been the first study to analyze BOSS spectra, since galaxy clustering studies like White et al. (2011) only use the spectra to confirm the nature of the object and obtain a redshift estimate.

To test our data reduction pipeline and our analysis methods we created synthetic datasets using the method described in chapter 2. The synthetic quasars were placed in the same angular and redshift position than the observed ones, and the spectra were normalized to match the observed magnitudes.

Thirty realizations of the whole dataset where generated, with different noise properties, quasar continua and with the option to include high column density systems and metals correlated with the Lyα forest. The analysis code was improved until we were sure that we could recover the input theory in the mocks in the absence of any systematic effect.

In figure 3.4 we show the result of fitting 30 mock realization of the survey, with different systematics added. We can see that the input theory is recovered in the absence of any systematic (top-left). The error bars increase when observational noise and continuum fluctuations are added to the mock spectra, but no systematic shift is introduced (top-right). The continuum fluctuations were introduced following the Principal Component Analysis (PCA) from Suzuki et al. (2005)).

Another important result shown in the figure is that the presence of high column density systems (bottom-left) or metals (bottom-right) can introduce a systematic error in the bias parameters recovered. The low value of the β parameter measured in the study could be explained by the presence of this systematic effect in the data.

As explained in subsection 3.3.2 we forced every individual spectrum to have $< \delta_F >= 0$ in the forest. This technique erases any mode with wavelength larger than the typical size of the forest. An analytical expression was computed to correct the predicted theory when comparing with it the measurement. We used the mock spectra to confirm the validity of this analytical correction.

Finally, the mock catalogs were used to test the code developed to compute the covariance matrix of the measurements.

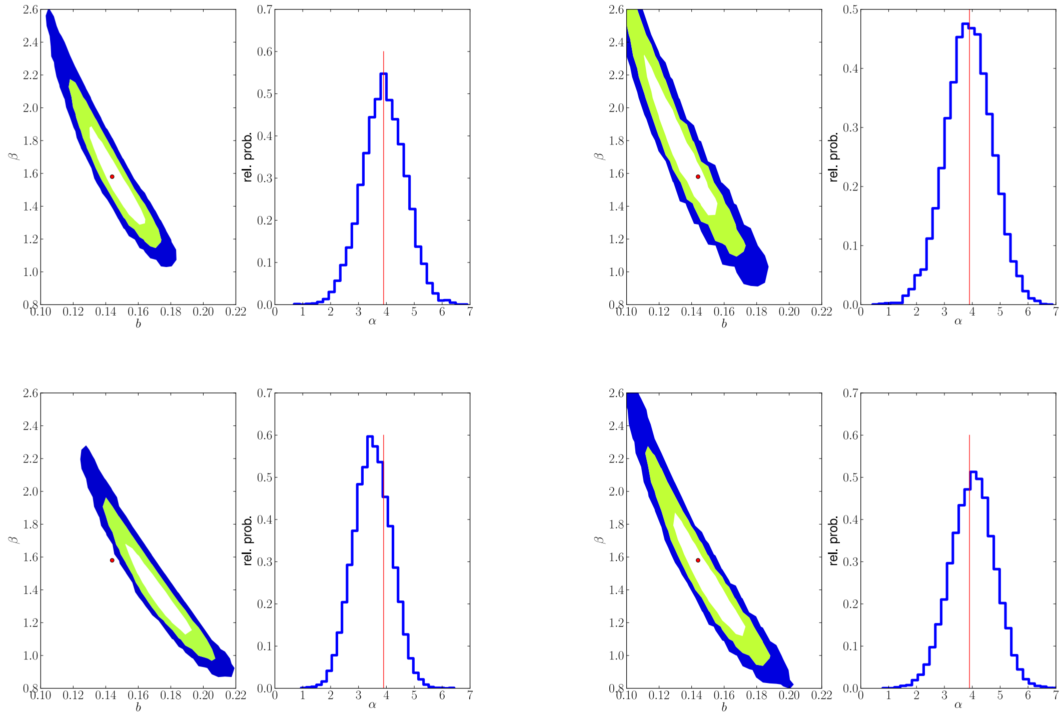

Figure 3.4: Results of fitting the data averaged over 30 mock datasets together with noise covariance for a single noisy realization and using only data points with $r > 20\,h^{-1}$ Mpc in the fit. We show constraints on the $b - \beta$ plane and the probability histogram of α (which has negligible degeneracy with the other parameters). The input points are denoted by the red dot and the red line. The upper left plot is for the pure synthetic noiseless δ_F values. The upper right plot is for synthetic data that have PCA continua and noise. The lower left plot is for the data that in addition to PCA continua are additionally painted with high column-density systems. The bottom right panel is for synthetic data to which metals have been added (with noise and continua but no hish column-density systems).

CHAPTER 4

The Effect of High Column Density Systems in the Measurement of the Lyman α Forest Correlation Function

4.1 Introduction

Observations of the correlation function of the Lyα forest in redshift space from multiple spectra is emerging as a powerful tool to explore the large-scale structure of the universe at high redshift. This development has been led by the BOSS survey, part of the SDSS-III collaboration (Eisenstein et al., 2011), which is obtaining optical spectra of 160 000 quasars at $z > 2.1$ for the main purpose of studying the Lyα forest absorption and measuring its power spectrum. The redshift space power spectrum of the fluctuations in the fraction of transmitted flux, F, can be quite complex on small scales (affected by non-linear gravitational evolution, thermal broadening, the non-linear relation between F and the optical depth...), but on large scales it should be simply related to the mass power spectrum in the linear regime, P_L, through two biasing parameters:

$$P_F(k, \mu_k) = b_\delta^2 (1 + \beta \mu_k^2)^2 \, P_L(k) \, , \tag{4.1.1}$$

where k and μ_k are the modulus and angle cosine relative to the line of sight of the wave vector in redshift space, b_δ is the bias factor relating the amplitude of fluctuations in F to the relative amplitude of density fluctuations, and β is the redshift distortion parameter. This form of the linear power spectrum in redshift space is the same as that for discrete tracers of the density field (Kaiser, 1987), except that β cannot be inferred from b_δ from the linear growth factor of density fluctuations, but is determined by a second independent bias parameter. Recently, the first measurement of b_δ and β for the Lyα forest was reported by Slosar et al. (2011) from the first year of BOSS data, and more accurate measurements are expected in the future.

The values of b_δ and β as a function of redshift can be predicted in principle from numerical simulations of the Lyα forest (McDonald, 2003, Slosar et al., 2009), and they depend on the detailed small-scale physical processes in the intergalactic medium. Comparison of the predicted values with the observed ones will therefore provide a means of testing these simulations of the evolving intergalactic medium. However, in practice the observed absorption spectra are affected not only by the low-density gas producing the Lyα forest, but also by higher density systems that give rise to absorption lines of high column density, observed as Lyman limit systems (LLS) and damped Lyα systems (DLA). These systems, as well as the lower column density Lyα forest, can additionally produce metal absorption lines, some of which appear in the region of the Lyα absorption, thereby contaminating the measurement of the Lyα power spectrum.

The presence of high column density systems (hereafter referred to as HCDs) have a similar effect on the Lyα power spectrum as the well-known "fingers of God" in galaxy redshift surveys: on small, non-linear scales, galaxies accumulate in high-density clusters with an internal velocity dispersion, which causes them to appear in redshift space as highly elongated structures along the line of sight. These "fingers of God" induce contours of the correlation function that are similarly elongated along the line of sight on small scales, precisely the opposite to the squashing effect on the correlation function contours induced by the $1 + \beta\mu_k^2$ term in the power spectrum that is prevalent on large, linear scales. In the case of absorption spectra, the damped wings of the HCDs have the same effect of spreading the correlation function along the line of sight, and metal lines with wavelengths overlapping the Lyα forest region can also have a similar effect. However, contrary to the "'fingers of God" in galaxy surveys, these effects extend out to all large scales in the Lyα forest, owing to the extension of damped wings and their substantial contribution to the total absorption, and the fact that metal lines may appear at any wavelength difference with respect to Lyα .

This paper focuses on the impact of HCDs on the bias factors of the Lyα forest. Their effect on the measured power spectrum is determined by the fact that HCDs are correlated with the underlying mass distribution and therefore with the Lyα forest intergalactic absorption. Their presense also adds additional noise to any power spectrum measurements. The impact of metal-line absorbers is also important and was briefly discussed in Slosar et al. (2011), but we shall not treat them in this paper.

The effects of HCDs on the Lyα absorption correlation function is discussed analytically in Section 4.2, and their effect is partially quantified in a specific model for their column density distribution in 4.3. A more complete calculation of the impact of HCDs is obtained through realistic mocks of Lyα spectra in 4.4.

A standard flat ΛCDM cosmology is used in this paper with the following parameters: $h = 0.72$, $\Omega_m = 0.281$, $\sigma_8 = 0.85$, $n_s = 0.963$, $\Omega_b = 0.0462$.

4.2 Analytical Description

The effect of High Column Density Systems (HCDs) on the Lyα correlation function can be partly described and computed analytically. We start this section introducing some useful notation. The transmitted fraction at a point $\mathbf{x}$ in the spectrum is $F(\mathbf{x}) = \bar{F}\,[1 + \delta_F(\mathbf{x})]$, where $\bar{F}$ is the mean value of F. In general, the absorption can be divided into contributions from the Lyα forest and from the HCDs, which must be conventionally separated at some column density. Here, we shall consider HCDs to be those systems with a column density that produces a continuous absorption optical depth greater than unity at the Lyman limit edge, $N_{HI} > 1.6 \times 10^{17}\,\mathrm{cm}^{-2}$. The important point is that the Lyα forest absorption is contributed mostly by systems with much lower column density, and the HCDs absorption is contributed by systems with much higher column density than this threshold, so the precise choice for the threshold is not crucial. These two absorption fields are designated as $F_\alpha(\mathbf{x}) = \bar{F}_\alpha\,[1+\delta_\alpha(\mathbf{x})]$, and $F_H(\mathbf{x}) = \bar{F}_H\,[1+\delta_H(\mathbf{x})]$, respectively. We then have

$$F(\mathbf{x}) = \bar{F}(1+\delta_F(\mathbf{x})) = F_\alpha(\mathbf{x})\,F_H(\mathbf{x}) = \bar{F}_\alpha\,[1+\delta_\alpha(\mathbf{x})]\,\bar{F}_H\,[1+\delta_H(\mathbf{x})]\ . \tag{4.2.1}$$

Because the fields δ_α and δ_H are tracers of the same underlying mass density field, they are correlated,

$$C \equiv \langle \delta_\alpha(\mathbf{x})\delta_H(\mathbf{x})\rangle \neq 0\,, \tag{4.2.2}$$

and the relation between $\bar{F}$ and $\bar{F}_\alpha$ is

$$\bar{F} = \langle F\rangle = \bar{F}_\alpha \bar{F}_H(1+C)\ . \tag{4.2.3}$$

Hence, the variable $\delta_F(\mathbf{x})$ can be expressed as

$$1+\delta_F(\mathbf{x}) = \frac{F(\mathbf{x})}{\bar{F}} = \frac{[1+\delta_\alpha(\mathbf{x})]\ [1+\delta_H(\mathbf{x})]}{1+C}\ . \tag{4.2.4}$$

4.2.1 Effect on the Correlation Function

We are now in a position to study the correlation function of δ_F between two points $\mathbf{x_1}$ and $\mathbf{x_2}$, with separation $\mathbf{r_{12}} = \mathbf{x_1} - \mathbf{x_2}$:

$$\begin{aligned} 1+\xi_F(\mathbf{r_{12}}) &= \langle [1+\delta_F(\mathbf{x_1})]\,[1+\delta_F(\mathbf{x_2})]\rangle \qquad (4.2.5)\\ &= (1+C)^{-2}\,\langle [1+\delta_\alpha(\mathbf{x_1})]\,[1+\delta_H(\mathbf{x_1})]\,[(1+\delta_\alpha(\mathbf{x_2})]\,[(1+\delta_H(\mathbf{x_2})]\rangle \\ &= (1+C)^{-2}\,[1+2C+\xi_\alpha(\mathbf{r_{12}})+2\xi_{\alpha H}(\mathbf{r_{12}})+\xi_H(\mathbf{r_{12}}) \\ &\quad +2\xi_{3\alpha}(\mathbf{r_{12}})+2\xi_{3H}(\mathbf{r_{12}})+\xi_4(\mathbf{r_{12}})] \end{aligned}$$

where we have defined:

$$\begin{aligned} \xi_\alpha(\mathbf{r_{12}}) &= \langle \delta_\alpha(\mathbf{x_1})\delta_\alpha(\mathbf{x_2})\rangle \\ \xi_{\alpha H}(\mathbf{r_{12}}) &= \langle \delta_\alpha(\mathbf{x_1})\delta_H(\mathbf{x_2})\rangle \\ \xi_H(\mathbf{r_{12}}) &= \langle \delta_H(\mathbf{x_1})\delta_H(\mathbf{x_2})\rangle \\ \xi_{3\alpha}(\mathbf{r_{12}}) &= \langle \delta_\alpha(\mathbf{x_1})\delta_H(\mathbf{x_1})\delta_\alpha(\mathbf{x_2})\rangle \\ \xi_{3H}(\mathbf{r_{12}}) &= \langle \delta_\alpha(\mathbf{x_1})\delta_H(\mathbf{x_1})\delta_H(\mathbf{x_2})\rangle \\ \xi_4(\mathbf{r_{12}}) &= \langle \delta_\alpha(\mathbf{x_1})\delta_H(\mathbf{x_1})\delta_\alpha(\mathbf{x_2})\delta_H(\mathbf{x_2})\rangle\ . \end{aligned} \tag{4.2.6}$$

If we now define an effective correlation function,

$$\xi_{eff}(\mathbf{r_{12}}) \equiv \frac{\xi_\alpha(\mathbf{r_{12}})+2\xi_{\alpha H}(\mathbf{r_{12}})+\xi_H(\mathbf{r_{12}})}{(1+C)^2}\,, \tag{4.2.7}$$

and include all the terms correlating three and four variables in the function

$$\xi_{34}(\mathbf{r_{12}}) \equiv \frac{2\xi_{3\alpha}(\mathbf{r_{12}}) + 2\xi_{3H}(\mathbf{r_{12}}) + \xi_4(\mathbf{r_{12}})}{(1+C)^2} , \tag{4.2.8}$$

equation 4.2.5 is compressed to:

$$\xi_F(\mathbf{r_{12}}) = \xi_{eff}(\mathbf{r_{12}}) + \xi_{34}(\mathbf{r_{12}}) - \left(\frac{C}{1+C}\right)^2 . \tag{4.2.9}$$

Note that in the limit of very large separation, the only non-vanishing term,

$$\frac{\xi_4(\mathbf{r_{12}})}{(1+C)^2} = \frac{C^2}{(1+C)^2} \tag{4.2.10}$$

cancels the constant term, yielding $\xi_F = 0$ as expected.

4.2.2 Effective Bias Parameters

The 3-point and 4-point functions we have defined in the previous section that affect the impact of the HCDs on the overall correlation function are difficult to characterize. However, the two-point correlation function terms are easily obtained from the fact that each Fourier mode is multiplied by the factor $1 + \beta_i \mu_k^2$ for any tracer with a redshift distortion factor β_i (Kaiser, 1987). The power spectra corresponding to ξ_α, ξ_H, and $\xi_{\alpha H}$ are

$$P_\alpha(k, \mu_k) = b_\alpha^2(1 + \beta_\alpha \mu_k^2)^2 P_L(k) , \tag{4.2.11}$$

$$P_H(k, \mu_k) = b_H^2(1 + \beta_H \mu_k^2)^2 P_L(k) , \tag{4.2.12}$$

$$P_{\alpha H}(k, \mu_k) = b_\alpha(1 + \beta_\alpha \mu_k^2) b_H (1 + \beta_H \mu_k^2) P_L(k) , \tag{4.2.13}$$

where $P_L(k)$ is the linear matter power spectrum, μ_k is the cosine of the angle of the wave vector relative to the line of sight, b_α and β_α are the usual bias parameters for the Lyα forest absorption, and b_H and β_H are the bias parameters of the HCD absorption.

The effective power spectrum (the Fourier transform of ξ_{eff}) can be expressed as

$$\frac{P_{eff}(k, \mu_k)}{P_L(k)} = \frac{b_\alpha^2(1 + \beta_\alpha \mu_k^2)^2 + 2 b_\alpha b_H (1 + \beta_\alpha \mu_k^2)(1 + \beta_H \mu_k^2) + b_H^2 (1 + \beta_H \mu_k^2)^2}{(1+C)^2} \tag{4.2.14}$$

$$= b_{eff}^2 (1 + \beta_{eff} \mu_k^2)^2 , \tag{4.2.15}$$

where

$$b_{eff} = \frac{b_\alpha + b_H}{1+C} , \tag{4.2.16}$$

and

$$b_{eff}\beta_{eff} = \frac{b_\alpha \beta_\alpha + b_H \beta_H}{1+C} . \tag{4.2.17}$$

In the absence of the term ξ_{34}, this result means that if we measure the Lyα forest bias parameters from the total transmitted fraction $F(x)$, the measurement will be systematically biased,

$$\Delta b \equiv b_{eff} - b_\alpha = \frac{b_H - b_\alpha C}{1 + C} \ , \tag{4.2.18}$$

and

$$\Delta \beta \equiv \beta_{eff} - \beta_\alpha = \frac{b_H(\beta_H - \beta_\alpha)}{b_\alpha + b_H} \ . \tag{4.2.19}$$

4.2.3 Relation to the Bias of Host Halos

Whereas most of the Lyα forest absorption at $z > 2$ is associated with density fluctuations in the intergalactic medium forming an interconnected structure, the high column density systems should correspond to discrete, clearly identifiable overdense regions that have gravitationally collapsed, or halos. Note that there is always some remaining ambiguity in the identification of halos as separate objects when the halos are in the process of merger events, but only a small fraction of halos are undergoing a merger at any given time. The question that arises then is the relation between the bias factor of the halos hosting the HCDs and the bias factor of the HCDs when measured in the absorption spectra. This relation is in general complicated because the HCDs in absorption are clustered and their absorption profiles in the spectra can be blended in a non-linear way, in which their absorption equivalent widths are not simply added up. However, the two bias factors should be simply related under the following two simplifying assumptions:

1. The probability that the absorption profile of any HCD appears substantially blended with another one in the absorption spectrum is small. Here, substantially blended means that their profiles overlap in a region where their absorption optical depth is close to or greater than unity. This condition should in general be correct if $1 - \bar{F}_H \ll 1$ and the clustering of HCDs is not very strong.

2. The probability distribution of the column density in a halo of a fixed mass M_h is independent of its large-scale environment and is isotropic. In other words, the axes of any non-spherical gas distribution in the halos are not aligned with the principal axes of the velocity gradient matrix in the surrounding large-scale structure. This assumption is likely to be not precisely true, because galaxy disks are known to be statistically aligned with the axes of their large-scale environment, and this can affect their redshift distortion anisotropy (Hirata, 2009), but the effect is probably small.

Under these conditions, the HCDs (whether they are individually detected or not) appear in the spectra following the same relative fluctuations as their host halos, with an overdensity field smoothed over large, linear scales, $\delta_h = b_h\delta$, where δ is the mass density perturbation and b_h is the halo bias factor. This results in a mean optical depth from HCDs equal to $\bar{\tau}_H(1+\delta_h)$ in any large region of halo overdensity δ_h, where $\bar{\tau}_H$ is the mean optical depth from HCDs in the entire universe. The transmission fluctuation is then given by

$$\delta_H(\mathbf{x}) = \log(\bar{F}_H)\delta_h(\mathbf{x}) \simeq -(1-\bar{F}_H)\delta_h(\mathbf{x}) \, , \tag{4.2.20}$$

where we use again $1-\bar{F}_H \ll 1$. Therefore, the bias factors are related by

$$b_H = -(1-\bar{F}_H)b_h \, ; \qquad\qquad \beta_H = \beta_h \, . \tag{4.2.21}$$

In this approximation, it is therefore true that, for HCDs, $\beta_H = f(\Omega)[(\bar{F}_H - 1)/b_H]$, where $f(\Omega)$ is the logarithmic derivative of the gravitational growth factor of linear perturbations (Kaiser, 1987). Note that a relation of this type does not exist for the Lyα forest, because the conditions mentioned previously are not correct.

4.3 Effective Bias in a Lyα Survey

We now quantify the effective bias parameters for the two-point component of the correlation function ξ_{eff} described in the previous section, for a specific model of the column density distribution of HCDs. We first compute the expected fraction of flux that is absorbed by HCDs in the redshift range of interest, and then we quantify the change induced in the bias parameters using the expressions derived in the previous section.

4.3.1 Column Density Distribution

The abundance of HCD systems is not easy to quantified from simulations or observations. The column density of the systems is determined by its H_I density, that is very sensitive to complex astrophysical processes difficult to simulate in numerical simulations.

The study is also challenging from the observational point of view, especially in the lower column density regime where we need high resolution spectra to resolve the systems and measure its equivalent width, at the same time that we need a large number of spectra to have a large enough statistics. On the other hand, DLAs have broader absorption profiles and are easier to detect even with the mid-resolution SDSS spectrograph (Prochaska et al., 2005).

Here we assume a neutral hydrogen column density distribution from McDonald et al. (2005), based on an analytical expression derived in Zheng and Miralda-Escudé (2002), and calibrated to match the observations of DLAs of Prochaska et al. (2005).

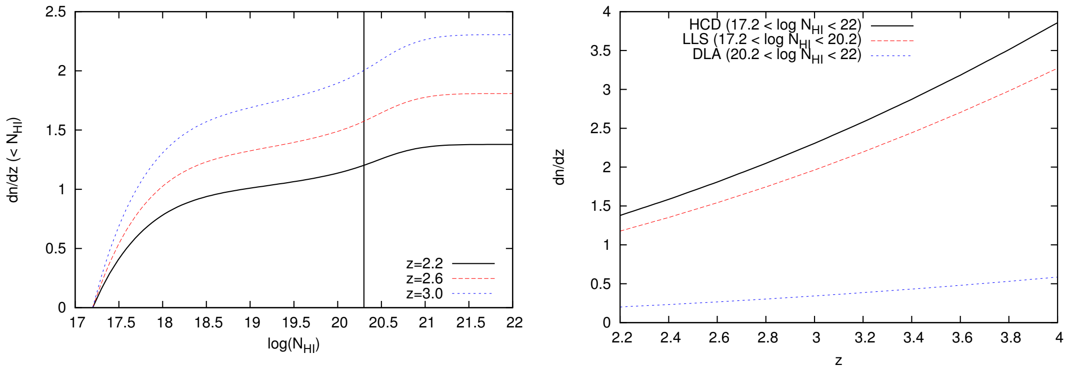

Figure 4.1: Left: Number of systems below column density N_{HI} per unit redshift, at three redshifts. The vertical line indicates the standard separation between DLA and LLS. Right: Number of systems (HCD, LLS and DLA) as a function of redshift.

In figure 4.1 we show the column density distribution at three different redshifts (left). The distribution flattens in the column density interval between 10^{18} and $10^{20}\,\mathrm{cm}^{-2}$ owing to the self-shielding of the ionizing radiation. In figure 4.1 we also plot the number of these systems (HCD, LLS and DLA) as a function of redshift (right).

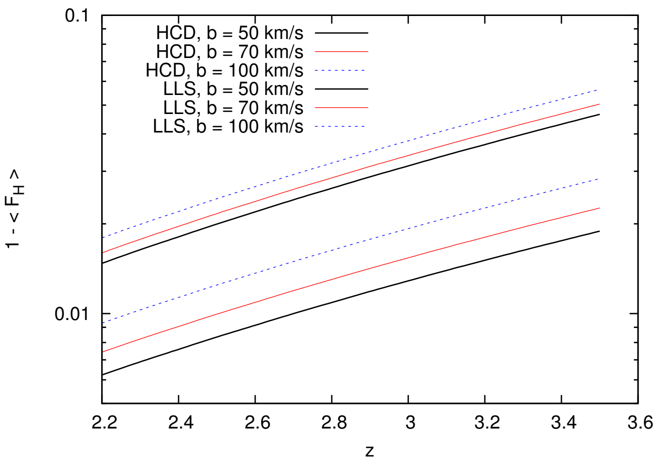

Figure 4.2: Mean absorption caused by HCD systems, for different values of b. The 3 lower lines show the effect due to LLS exclusively.

The absorption feature caused by a HCD system is not determined by its column density N_{HI} but also depends on the velocity dispersion (or the temperature) of the gas in the system, quantified with the Doppler parameter b. This parameter enters also in the curve of growth, the relation between the equivalent with W of a system and its column density N_{HI}, and effects the fraction $\bar{F_H}$ of flux absorbed by HCD systems:

$$\bar{F_H}(z) = \int dN_{HI} \, \frac{dn(z, N_{HI})}{dN_{HI}dz} \, \frac{W(N_{HI}, b)}{\lambda_\alpha} \, (1+z) \; . \qquad (4.3.1)$$

In figure 4.2 we plot $1 - \bar{F_H}$ as a function of redshift, distinguishing between the fraction absorbed by DLAs and LLS, for typical values of b. [1] We can see that the value of $\bar{F_H}$ evolves from $\bar{F_H} \sim 0.95$ at $z = 3.5$ to $\bar{F_H} \sim 0.98$ at $z = 2.5$ if considering all HCD systems, while the effect caused by LLS evolves from $\bar{F_H} \sim 0.98$ at $z = 3.5$ to $\bar{F_H} \sim 0.99$ at $z = 2.5$

4.3.2 Bias of the Host Halos

Since the nature of the HCD systems is not well known, it is difficult to estimate the bias of the host halos. Detailed cosmological simulations (Pontzen et al., 2008) show that a typical mass for the halos hosting DLAs is $10^9 - 10^{11} \mathrm{M}_\odot$, but its not clear that this result also applies for LLSs. Using a simple mass-bias relation (Press and Schechter, 1974, Sheth and Tormen, 1999) we can translate this result into a range for the typical bias for the host halo of $1 < b_h < 2$.

As suggested in Chapter 5, we should be able to measure the bias of halos hosting DLAs with a $\sim 30\%$ accuracy via their cross-correlation with the Lyα forest with the BOSS survey, and use this result to better understand their effect in the Lyα correlation function.

Finally, The value of C depends on the details of how HCDs trace the density field, and hence a detailed simulation of both the systems and the Lyα absorption is needed to accurately compute its value. Using the mocks explained in Section 4.4 we find a value of $C \sim 6 \times 10^{-3}$. This value should be taken as a rough approximation, but fortunately the effective bias parameter b_{eff} is only weakly dependent on the C parameter (as long as it is small) and the redshift space parameter β_{eff} is completely independent.

4.3.3 Expected Values

We can now compute the expected effective biases for different values of the unknown parameters:

In table 4.1 we show the systematic effect in the inferred Lyα forest bias parameters due

[1] In the mocks described in the next section, we use a value of $b = 70 km/s$.

Δb		b_h		
		1.0	1.5	2.0
$\bar{F_H}$	0.99	-0.0094	-0.014	-0.019
	0.98	-0.020	-0.030	-0.040
	0.95	-0.052	-0.078	-0.10

$\Delta\beta$		b_h		
		1.0	1.5	2.0
$\bar{F_H}$	0.99	-0.042	-0.095	-0.15
	0.98	-0.079	-0.17	-0.26
	0.95	-0.17	-0.35	-0.48

Table 4.1: Systematic effect in the Lyα forest bias parameters due to the presence of HCDs, ignoring the ξ_{34} term. The other parameters are fixed to $b_\alpha = -0.1315$, $\beta_\alpha = 1.58$, $C = 0.006$, and we consider $\beta_h = b_h^{-1}$.

to the term ξ_{eff}, for different values of $\bar{F_H}$, b_h, assuming $\beta_g = b_h^{-1}$ and fixing the other parameters [2] to $b_\alpha = -0.1315$, $\beta_\alpha = 1.58$ and $C = 0.006$.

We can see in the table that the allowed range of values for Δb, $\Delta\beta$ is very large. For the reasonable values of $\bar{F_H} = 0.98$ and $b_h = 1.5$, the relative effects are:

$$\frac{\Delta b}{b_\alpha} = 0.23, \qquad \frac{\Delta\beta}{\beta_\alpha} = 0.11 \ . \tag{4.3.2}$$

In a survey with similar characteristics as BOSS one should be able to detect and mask the systems with higher column density, specially in high S/N spectra, reducing considerably the value of $\bar{F_H}$. In figure 4.2 we also plot the value of $\bar{F_H}$ as a function of redshift for the case where only LSS are responsible for the absorption because all DLAs have been masked. As mentioned above, the detectability of the systems will be strongly effected by the S/N of the quasar, but this plot allows us to have a rough estimate of $\bar{F_H} = 0.99$ if most of the DLAs have been masked. In this case, and using also $b_h = 1.5$, the effect is considerably reduced:

$$\frac{\Delta b}{b_\alpha} = 0.11, \qquad \frac{\Delta\beta}{\beta_\alpha} = 0.06 \ . \tag{4.3.3}$$

4.4 Effect of High Column Density Systems in Mock Spectra

In the previous section we have quantified the expected change in the bias parameters caused by the presence of HCD systems caused by the term ξ_{eff}. In order to quantify the effect of the other term ξ_{34} detailed simulations of both the Lyα fores and HCD systems are needed.

Here we explain a method to introduce HCD systems to mock Lyα spectra, and we apply it to the mock spectra explained in Chapter 2. We then measure the correlation

[2] central values from McDonald (2003).

function of the mock spectra before and after adding the systems, and study the change in the measure of the bias parameters.

4.4.1 Lyα Mock Spectra

In Chapter 2 we develop a method to generate Lyα mock spectra with any given flux power spectrum and any probability distribution function of the flux. We refer the reader to that chapter and highlight here its main features.

The method consists in two steps:

- We generate a Gaussian random field $\delta_g(x)$ for a given set of correlated lines of sight.
- We apply a transformation $F(\delta_g)$ to a variable constrained in the range $0 < F < 1$. The power spectrum for the Gaussian variable is chosen in order to obtain the desired flux power spectrum after the transformation.

In Chapter 2 we apply a third step where we interpolate between lines of sight generated at different redshifts to simulate the effect of redshift evolution and to take into account the fact that the lines of sight are not parallel. In this paper we use a simplified version where the lines of sight have been generated at a fixed redshift of $z = 2.6$. To make easier the fitting of the bias parameters explained below, we use the simple linear theory power spectrum [3] described in Section 4.2,

$$P_\alpha(k,\mu) = b_\alpha^2 \, (1+\beta_\alpha\mu^2)^2 \, P_L(k) \; . \tag{4.4.1}$$

We use again the central values from McDonald (2003) for β, but we increase the value of b_α to take into account the evolution of the amplitude of the power spectrum with redshift $(1+z)^\alpha$, $\alpha = 3.8$, observed in the measurement of the 1D power spectrum of McDonald et al. (2006). Since we are generating the field at a redshift $z = 2.6$ instead of the central redshift $z_c = 2.25$ where the bias were calibrated, we have to apply the correction

$$b_\alpha(z) = b_\alpha(z_c) \, \left(\frac{1+z}{1+z_c}\right)^{\alpha/2} \, \frac{G(z_c)}{G(z)} \; , \tag{4.4.2}$$

finding a value of $b_\alpha(z = 2.6) = -0.177$.

[3] i.e. we do not add the small scale correction $D(k,\mu)$ from McDonald (2003) that was used in Chapter 2.

4.4.2 Adding Correlated Systems to the Lyα Mocks

Using the column density distribution explained in the previous section we could randomly introduce systems in our spectra to study the impact in the errorbars of our measurements. But if we want to study how they bias the measurement we need the systems to be correlated with the Lyα absorption. Here we present a simple method to place these systems in the peaks of the Lyα absorption field and we study its effect on the recovered Lyα forest statistics in the next subsection.

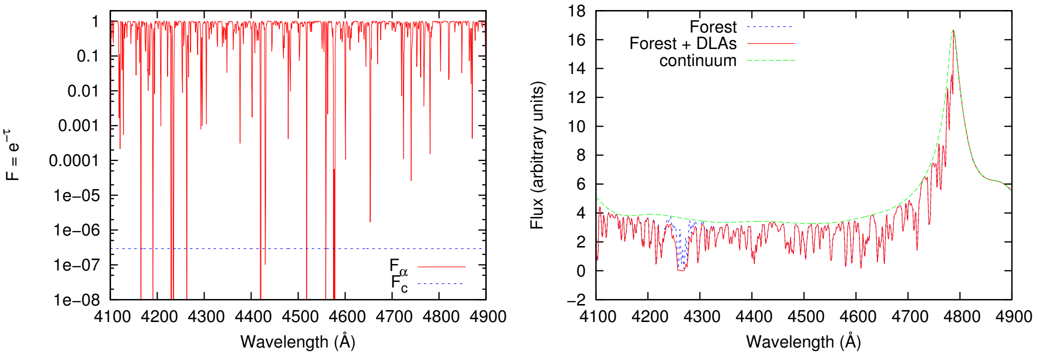

Figure 4.3: Left: Example of mock Lyα absorption field F (red) and the threshold F_c to host a HCD for $\nu = 0.01$ (blue). Right: Mock spectrum for the same line of sight. The green line shows a typical continuum for a quasar. The blue line includes the absorption due to Lyα forest, smoothed with the spectrograph resolution as explained in Chapter 2. The red line adds on top a HCD.

The method consists in placing the systems only in the fraction ν of pixels with higher optical depth. The value of ν (we use $\nu = 0.01$ here) determines a critical optical depth τ_c or, equivalently, a critical transmitted flux fraction F_c:

$$\nu = \int_{\tau_c}^{\infty} d\tau p_\tau(\tau) = \int_0^{F_c} dF p_F(F) \tag{4.4.3}$$

Because the probability distribution of optical depth $p_\tau(\tau)$ and transmitted flux fraction $p_F(F)$ depend on redshift, in general the threshold for hosting a HCD will also depend on redshift (for a fixed value of ν). Here we generate the survey at a fixed redshift so the value of F_c will be fixed.

Once we have identified the candidate pixels we distribute the systems with the column density distribution described in the previous section. [4] In figure 4.3 we show

[4] In Slosar et al. (2011) the systems were introduced at constant abundance in comoving separation (calibrated at $z = 2.6$), causing an overabundance of systems at low redshift and an underabundance at high redshift.

a typical line of sight (red) and the value of F_c used here (green). In the second figure, we show a mock spectra for the same line of sight, where the absorption field has been multiplied by a typical quasar continuum and smoothed with the resolution of the BOSS spectrograph as explained in Chapter 2. A DLA is randomly assigned to one of the peaks that cross the threshold in the first figure (for instance, the peak at $\lambda \sim 4260$ Å), and is also included to the absorption field in the green line of the second figure.

Bias of the HCD systems

The value of ν determines the level of clustering of the systems added. In appendix A.2 we show that the peaks of a Gaussian field have, on large scales, the same correlation function than the Gaussian field δ_g itself, but with a relative bias b_h/b_g given by:

$$\left(\frac{b_h}{b_g}\right)^2 = \frac{p_g(\delta_{gc})}{\nu^2} \int_{\delta_{gc}}^{\infty} d\delta_g \; p_g(\delta_g) \; \delta_g \; , \tag{4.4.4}$$

where b_g is the bias of the Gaussian field, δ_{gc} is the threshold to host a system and $p_g(\delta_g)$ is the Gaussian probability distribution.

The correlation of the Gaussian field itself is at the same time proportional to the Lyα forest correlation on large scales (as shown in appendix A.1):

$$\left(\frac{b_\alpha}{b_g}\right)^2 = \frac{1}{\bar{F}^2} \int_{-\infty}^{\infty} d\delta_g \; p_g(\delta_g) \; F(\delta_g) \; \delta_g \int_{-\infty}^{\infty} d\delta_g \; p_g(\delta_g) \; F(\delta_g) \; \frac{d\tau}{\delta_g} \; . \tag{4.4.5}$$

As a result, the HCD systems added to the mock spectra share the same correlation function than the Lyα forest field, with a different bias parameter, but with the same redshift distortions parameter β. The bias parameter of the systems depends not only on the ν parameter, but also in the transformation $F(\delta_g)$ used. Here we use a lognormal transformation for the optical depth τ

$$F = e^{-\tau} = e^{-ae^{\gamma\delta_g}} \tag{4.4.6}$$

with $a = 0.1566$, and $\gamma = 1.761$. Using this transformation and $\nu = 0.01$ the bias of the systems at the redshift used $z = 2.6$ is $b_h = 1.3$, where we have defined the bias as:

$$P_h(k,\mu) = b_h^2 \left(1 + \beta_h \mu^2\right)^2 P_L(k) \; . \tag{4.4.7}$$

4.4.3 Effect on the Measured Correlation Function

Following Chapter 2 we generate 50 realizations of a mock survey of $A = 300\,\mathrm{deg}^2$, covering a redshift range of $2.15 < z < 3.5$. We distribute quasar following the luminosity function from Jiang et al. (2006), but discard 25% of them to obtain a quasar density of $15 - 17\,\mathrm{deg}^{-2}$, similar to the observed during the first year of BOSS observations. As explained before, we do not introduce redshift evolution and generate the field at a fixed redshift of $z = 2.6$ instead.

For each realization we introduce HCD systems with the method explained above, and measure the correlation function in thin bins of r, $\mu = cos(\theta)$ (20 bins in μ and 150 bins in r of $1\,h^{-1}$ Mpc). We estimate the correlation function of a bin A by averaging the product of all pixel pairs pixels with a separation r and angle μ that are within a bin A:

$$\hat{\xi}_A = \frac{\sum_{i,j\in A} \delta_{Fi}\delta_{Fj}}{\sum_{i,j}} \,. \tag{4.4.8}$$

Here we do not weight our pixels since the spectra are noiseless and there is no redshift evolution either.

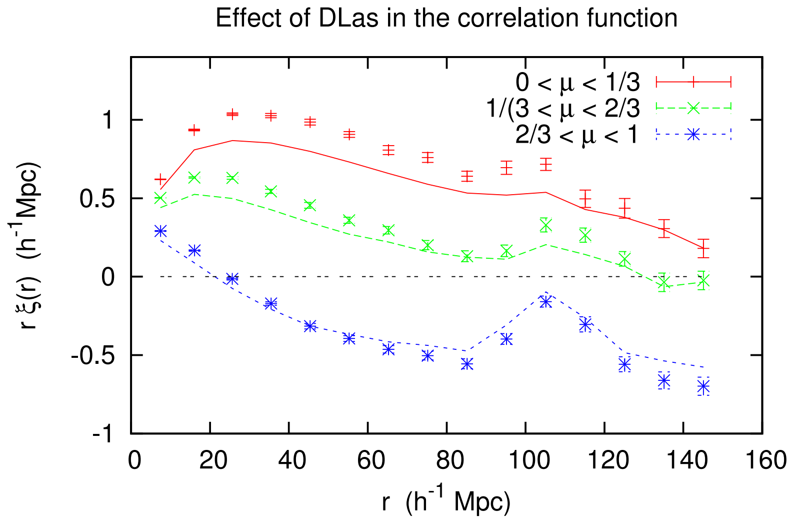

Figure 4.4: Correlation function measured from mock spectra with high column density systems added (points with errorbars) in 3 angular bins. The lines show the mean measured in the same mock spectra, but without the systems. The errorbars here are the errors in the mean, i.e. the dispersion divided by $\sqrt{50-1} = 7$.

In figure 4.4 we show the mean measurement of the correlation function in the 50 realizations, before and after adding the HCD systems, where the narrow bins have been compressed into bins of $\Delta r = 10\,h^{-1}$ Mpc and 3 μ bins. The errorbars are the error in the mean value, i.e. the dispersion between realizations divided by the squared root of the number of realizations minus one $\sqrt{50-1} = 7$. Once can see by eye that the presence of these systems increases the amplitude of the correlation. We quantify this effect in the next section.

4.4.4 Effect on the Bias Parameters

Following Yoo et al. (2009) and Slosar et al. (2011) we decompose the redshift space correlation function $\xi_F(r,\mu)$ into multipoles:

$$\xi_F(r,\mu) = b_F^2 \sum_{l=0,2,4} L_l(\mu)\; K_l(\beta_F)\; \xi_l(r)\;, \tag{4.4.9}$$

where $L_l(x)$ are the Legendre polynomials and we have defined the parameters:

$$K_0(\beta_F) = 1 + \frac{2}{3}\beta_F + \frac{1}{5}\beta_F^2\;, \qquad K_2(\beta_F) = \frac{4}{3}\beta_F + \frac{4}{7}\beta_F^2\;, \qquad K_4(\beta_F) = \frac{8}{35}\beta_F^2\;, \tag{4.4.10}$$

ξ_l are computed from:

$$\xi_l(r) = \frac{i^{-l}}{2\pi^2}\int_0^\infty dk\, k^2\, j_l(kr)\; P_L(k), \tag{4.4.11}$$

and $j_l(x)$ is the spherical Bessel function of order l.

We can now measure the Lyα forest multipoles ξ_{lF} from any data set (or a mock catalog) as follow:

$$\xi_{lF}(r) = \frac{2l+1}{2}\int_{-1}^{1} d\mu\; \xi_F(r,\mu)\; L_l(\mu)\;, \tag{4.4.12}$$

and using the equations above we can relate them to the functions ξ_l:

$$\xi_{lF}(r) = b_F^2\; K_l(\beta_F)\; \xi_l(r)\;. \tag{4.4.13}$$

In figure 4.5 we plot the multipoles measured from the mock catalogs, with (blue) and without (red) HCD systems added. The black lines show the expected values using the equations above.

Note that the multiplicative factors in 4.4.13 that relate the measured multipoles $\xi_{lF}(r)$ and the theoretical functions $\xi_l(r)$ are scale independent. This allows us to compress all the radial information and define a new set of variables X_l that are function of b_F and β_F:

$$X_l \equiv \frac{\int_{r_1}^{r_2} dr\; \xi_{Fl}(r)\; w(r)}{\int_{r_1}^{r_2} dr\; \xi_l(r)\; w(r)} = b_F^2\; K_l(\beta_F)\;, \tag{4.4.14}$$

where $w(r)$ is a weight to optimize the signal (we use the inverse of the variance between realizations) and we use $r_1 = 10\,h^{-1}$ Mpc and $r_2 = 80\,h^{-1}$ Mpc.

From the 50 realizations explained above, we can compute the main value of $\hat{X}_i =< X_i >$ and the elements of the covariance matrix C_{ij}:

$$C_{ij} =< X_i X_j > - \hat{X}_i \hat{X}_j\;. \tag{4.4.15}$$

We can now compute the likelihood function L for any set of parameters b_F, β_F, assuming Gaussian errors:

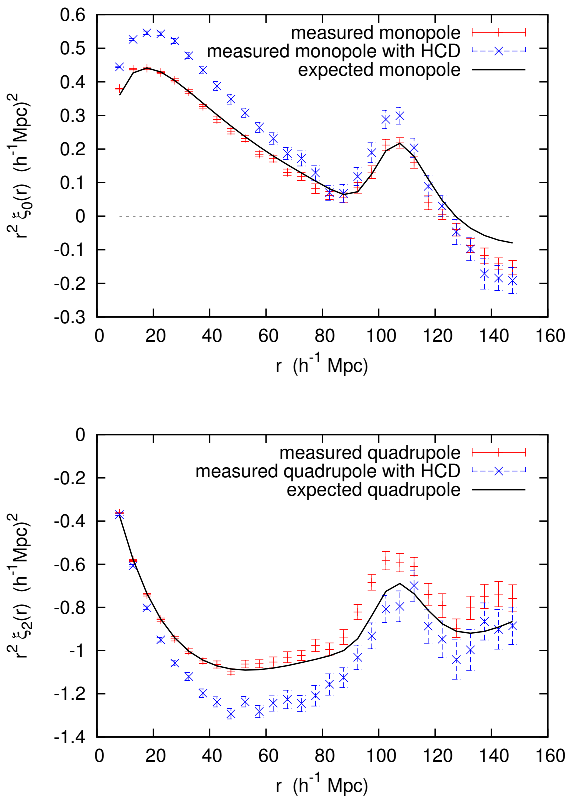

Figure 4.5: Multipoles of the correlation function measured from mock spectra with (blue) and without (red) HCD systems. The errorbars are again the errors in the mean, i.e. the dispersion divided by $\sqrt{50-1} = 7$. The input theory is also plotted (black line).

$$L(b_F, \beta_F) = \frac{e^{-\frac{1}{2}\,\chi^2(b_F,\beta_F)}}{2\pi\,|C|^{1/2}}\,, \tag{4.4.16}$$

with

$$\chi^2(b_F, \beta_F) = (\hat{X}_i - b_F^2\,K_i(\beta_F))\,C_{ij}^{-1}\,(\hat{X}_j - b_F^2\,K_j(\beta_F))\,. \tag{4.4.17}$$

In figure 4.6 we show the $1, 2, 3 - \sigma$ contours of the $b_F - \beta_F$ plane that best fit the measured X_l from mocks, before (left) and after (right) adding the HCD systems. Since the contours are highly degenerated we also plot the contours in the $b_F - b_f(1+\beta_F)$ plane. The shift in the best fit parameters due to the presence of HCD is:

$$\Delta b = -0.0373, \qquad \Delta\beta = -0.261\,. \tag{4.4.18}$$

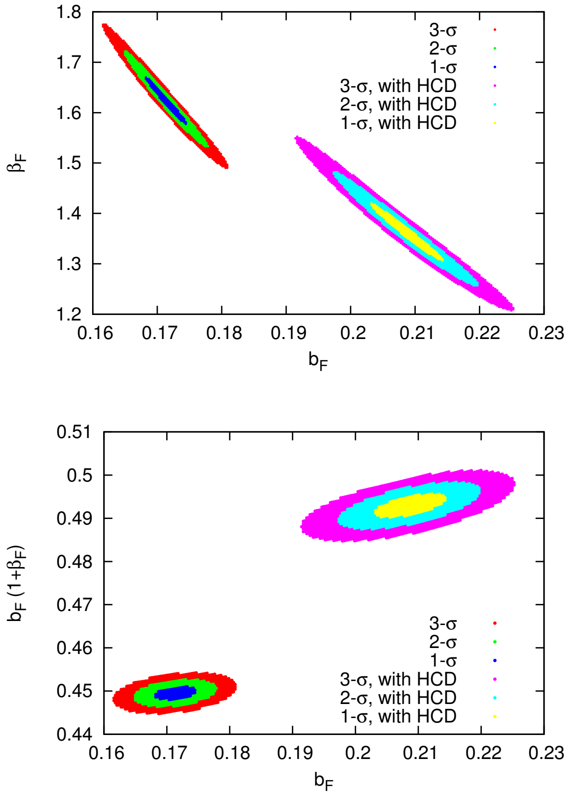

Figure 4.6: $1, 2, 3 - \sigma$ contours of the b_F - β_F plane fitted from the measured multipoles, without (left) and with (right) HCD systems.

Effect of $\xi_{>34}$ on β_F

The effective bias parameters that predict equations 4.2.16, 4.2.17 can be computed for the values used here: $b_h = 1.3$, $\beta_h = 1.58$, $C = 0.006$, $\beta_F = 1.58$, $b_F = -0.177$ and $\bar{F_H} = 0.98$:

$$\Delta b = -0.0269, \qquad \Delta\beta = 0\,. \tag{4.4.19}$$

Note that according to equations 4.2.17 the value of β_F should not change in our mocks because the systems have been introduced with the same redshift distortions ($\beta_H = \beta_\alpha$), although this is only true on large scales.

The discrepancies between the predicted change in the bias parameters and the measured bias could be caused by the ξ_{34} term that is not taken into account when computing the effective parameters.

This result suggest that not only ξ_{eff} lowers the value of the β_F parameter, but also ξ_{34} can substantially reduce it. This result should be confirmed by numerical simulations of both the HCD systems and the Lyα forest.

4.5 Conclusions

We have developed a formalism to analytically describe the impact of HCD systems in the measurement of the Lyα correlation function and in the inferred linear bias parameters. We divide the effect in two different terms:

- ξ_{eff} groups the contributions from 2-points correlation functions of both the HCD systems and the Lyα forest. On large scales is described with effective bias parameters b_{eff}, β_{eff}.
- ξ_{34} groups the contributions from 3 and 4-points correlations and can not be computed analytically.

In Section 4.3 we compute the effective bias parameters that one measure in a Lyα survey as BOSS. For typical values of the HCD bias parameter and abundances, the effects are a $\sim$ 10% decrease in the β_F parameter and a a $\sim$ 20% increase in b_F (in absolute value).

In Section 4.4 we have measured the effect in mock catalogs where HCD have been added, and find that the term ξ_{34} is non-negligible and it could also lower the inferred value of β_F.

Even though some of the larger systems could be systematically masked in Lyα surveys like BOSS, the smaller systems would still effect the measurement of the correlation function and this effect should be taken into account when interpreting the results.

CHAPTER 5

Crosscorrelating the Lyman-α Forest

5.1 Introduction

In the current models of galaxy formation, galaxies are formed in Dark Matter (DM) halos, gravitationally collapsed objects arising from originally small density fluctuations in the early universe. Galaxies and halos are not randomly distributed in the Universe, but trace the underlying large scale density fluctuations. Massive galaxy surveys confirm large-scale structure in their distribution, and the existance of large galaxy clusters, voids and filamentary structure.

All objects do not trace the matter density in the same way. The relative strength between the clustering of the objects and that of the underlying matter is, on large scales, quantified by the linear bias parameter, b. It is possible to derive simple relations between the mass of a collapsed object (or halo) and the value of its bias parameter, at different epochs. Massive objects are rare and tend to appear only in large overdense regions of the universe, so their clustering strength is higher. The growth of structure with time makes a gravitationally bound object to be less rare as the universe evolves, reducing its bias parameter with time.

A measurement of the bias parameter of a collapsed object provides valuble information on the nature of the object, and the typical mass of the halos that host them. As an example, observations from galaxy surveys show that red galaxies are more clustered than blue galaxies, indicating that red galaxies tend to live in more massive halos.

There are two main difficulties in measuring the clustering of galaxies using a survey. On one hand, the volume of the survey limits the number of Fourier modes available, and hence the result may be affected by statistical fluctuations, also known as cosmic variance. On the other hand, the fact that we are sampling the Fourier modes with only a finite number of points introduces a second uncertainty, known as "shot noise", that is inversely proportional to the number density of objects.

When studying the clustering of rare objects (such as DLAs, MgII or CIV absorbers, etc.) using large spectroscopic surveys, it is clear that the latter uncertainty dominates, since the number of Fourier modes is large and the density of objects is low.

The crosscorrelation of these rare systems with common galaxies has been proposed to study their clustering in low redshifts, where the clustering of galaxies is well known. In the redshift range $2 < z < 3$, galaxies are difficult to detect, making them unsuitable for cross correlations. However, at this redshift range we can correlate the objects with the absorption features present in quasar spectra that are caused by neutral hydrogen absorption, namely the Lyα forest.

There are several interesting candidates to be correlated with the Lyα forest. Quasars are abundant enough to measure their autocorrelation ,but the crosscorrelation might provide an independent measurement with different systematic uncertainties, with comparable accuracy. Metal absorption systems in the same quasar spectra can also be correlated with the Lyα forest, providing information on the galaxies associated to the metals. Finally, Damped Lyα systems (DLAs) can also be detected in quasar spectra and are also candidates to correlate with the Lyα forest.

The systems causing the DLAs are thought to host most of the neutral hydrogen at these redshifts, but their nature is still unclear, even though tailed numerical simulations (Pontzen et al., 2008) suggest that they are the hosted in galaxies with mass in the range $10^9 - 10^{11}\mathrm{M}_\odot$. Wyithe (2008) proposed to measure the mass of the halos hosting DLAs using the fluctuations in the 21-cm emission. The expected uncertainties from a second generation of low-frequency arrays are of the order few tens of percent.

We start this study by presenting a simple Fisher matrix forecast to estimate the expected uncertainties in the crosscorrelation of a generic collapsed object, (galaxy or other system) with the Lyα forest in a spectroscopic survey with properties similar to those of the Baryon Oscillation Spectroscopic Survey (BOSS) (Eisenstein et al., 2011). We continue by presenting a simple method to compute the crosscorrelation by "stacking the flux around the objects". This technique does not require any knowledge of the selection function of the objects, and allows a simple estimation of the covariance matrix of the bins measured. Finally, we apply this method to measure the DLA-Lyα cross correlation in mock spectra and compute the expected uncertainty in the bias parameter that one could obtain from BOSS.

During this study, we consider collapsed objects as galaxies.

A standard flat ΛCDM cosmology is used in this paper with the following parameters: $h = 0.72$, $\Omega_m = 0.281$, $\sigma_8 = 0.85$, $n_s = 0.963$, $\Omega_b = 0.0462$.

5.2 Fisher Matrix Forecasts

Here we compute a Fisher matrix forecast of the signal to noise (S/N) that one can achieve in the measurement of the crosscorrelation of Lyα absorption with any kind of galaxies.

We first compute what is the S/N for the measurement of the cross power spectrum in a given bin in k, and compare it with the S/N for the autocorrelation of both galaxies and Lyα forest. We then compute the expected uncertainties in the measurement of the galaxy bias parameters, using a Fisher matrix approach.

5.2.1 Signal to Noise Estimation

Here we compute the S/N for the crosscorrelation that one could obtain from a spectroscopic survey, and compare it with the signal to noise ratio for the autocorrelation of galaxies and of Lyα forest.

Galaxy autocorrelation

The number of galaxies at a given position, $g(\mathbf{x}) = \bar{g}\left[1 + \delta_g(\mathbf{x})\right]$ is related to the underlying matter density $\delta_m(\mathbf{x})$. On large scales we can use the linear bias model that relates both Fourier modes,

$$\tilde{\delta}_g(\mathbf{k}) = b_g\left[1 + \beta_g \mu_k^2\right] \tilde{\delta}_m(\mathbf{k})\;, \tag{5.2.1}$$

where b_g is the galaxy bias, μ_k is the cosine of the angle of $\mathbf{k}$ relative to the line of sight and $\beta_g = f(\Omega_m)/b_g$ is the redshift distortion parameter [1] , with $f(\Omega_m)$ being the growth of structure rate that depends on the matter density parameter Ω_m.

The amplitude of the Fourier modes is given by the galaxy power spectrum

$$\left\langle \tilde{\delta}_g(\mathbf{k})\tilde{\delta}_g^{\star}(\mathbf{k\prime})\right\rangle = (2\pi)^3\delta^D(\mathbf{k}-\mathbf{k\prime})P_g(\mathbf{k})\;, \tag{5.2.2}$$

and in the linear regime is related to the linear matter power spectrum $P_L(k)$ (Kaiser, 1987),

$$P_g(\mathbf{k}) = b_g^2\left[1 + \beta_g \mu_k^2\right]^2 P_L(k)\;. \tag{5.2.3}$$

The accuracy with what one can measure the galaxy power spectrum $P_g(\mathbf{k})$ in a given bin in ($k \pm \Delta k$,$\mu_k \pm \Delta\mu_k$) can be quantified by the signal to noise ratio (S/N),

$$\left(\frac{S}{N}\right)_g^2 = N_k\,\frac{P_g^2(k,\mu_k)}{var[P_g(k,\mu_k)]}\;, \tag{5.2.4}$$

where N_k is the number of modes in the bin,

$$N_k = \frac{k^2\;\Delta k\;\Delta\mu_k\;A\;L}{2\pi^2}\;, \tag{5.2.5}$$

and A and L are the survey area and depth.

[1] This relation is only valid for objects whose selection function does not depend on the velocity gradient. This is clearly not the case for DLA, since their detectability depends on the $\delta_F(\mathbf{x})$ at the pixel, and this depends on the velocity gradient. We ignore this problem in this study.

The variance in the measured power spectrum can be approximated by

$$var\left[P_g(k,\mu_k)\right] = 2\left(P_g(k,\mu_k) + n_g^{-1}\right)^2 , \tag{5.2.6}$$

with n_g the number density of objects.

The S/N can then be described as

$$\left(\frac{S}{N}\right)_g^2 = N_k \,\frac{P_g^2(k,\mu_k)}{2\left(P_g(k,\mu_k) + n_g^{-1}\right)^2} . \tag{5.2.7}$$

Lyα autocorrelation

The Lyα absorption is usually quantified by the transmitted flux fraction $F(\mathbf{x}) = exp\left[-\tau(\mathbf{x})\right]$, where $\tau(\mathbf{x})$ is the optical depth. The fluctuations around the mean value $\delta_F(\mathbf{x}) = F(\mathbf{x})/\bar{F} - 1$ are also related to the underlying matter density. Again, the linear bias model relates both Fourier modes on large scales,

$$\tilde{\delta}_F(\mathbf{k}) = b_F\left[1+\beta_F\mu_k^2\right]\tilde{\delta}_m(\mathbf{k}) , \tag{5.2.8}$$

where b_F is the Lyα bias and now β_F has to be treated as a free parameter.

The amplitude of the Fourier modes is given by the Lyα power spectrum, and again it can be related to the linear matter power spectrum on large scales,

$$\langle\tilde{\delta}_F(\mathbf{k})\,\tilde{\delta}_F^{\star}(\mathbf{k}\prime)\rangle = (2\pi)^3\,\delta^D(\mathbf{k}-\mathbf{k}\prime)\,P_F(\mathbf{k}) = (2\pi)^3\,\delta^D(\mathbf{k}-\mathbf{k}\prime)\,b_F^2\left[1+\beta_F\mu_k^2\right]^2 P_L(k) . \tag{5.2.9}$$

McDonald and Eisenstein (2007) computed the expected S/N in the measurement of $P_F(k,\mu_k)$ in a spectroscopic survey, and highlighted the importance of the " aliasing term " due to the sparse sampling of the universe. Here we use the formalism from McQuinn and White (2011) that combines both the noise term and the aliasing term defining a noise-weighted density of lines of sight per unit area n_{eff},

$$\left(\frac{S}{N}\right)_F^2 = N_k\,\frac{P_F^2(k,\mu_k)}{var[P_F(k,\mu_k)]} = N_k\,\frac{P_F^2(k,\mu_k)}{2\left(P_F(k,\mu_k) + P^{1D}(k\mu_k)\,n_{eff}^{-1}\right)^2} , \tag{5.2.10}$$

where $P^{1D}(k\mu_k)$ is the one-dimensional flux power spectrum.

Crosscorrelation

The cross correlation between the Lyα absorption and any galaxy field can be defined as

$$\left\langle\tilde{\delta}_F(\mathbf{k})\,\tilde{\delta}_g^{\star}(\mathbf{k}\prime)\right\rangle = (2\pi)^3\delta^D(\mathbf{k}-\mathbf{k}\prime)\,P_{gF}(k,\mu_k) . \tag{5.2.11}$$

Again, in the linear regime we can relate the crosscorrelation power spectrum with the linear power spectrum $P_L(k)$ using the linear bias parameters defined above,

$$P_{gF}(\mathbf{k}) = b_g \left[1 + \beta_g \mu_k^2\right] \; b_F \left[1 + \beta_F \mu_k^2\right] P_L(k) \, . \tag{5.2.12}$$

McQuinn and White (2011) showed that the variance in the measurement of the crosscorrelation can be approximated by

$$var\left[P_{gF}(k,\mu_k)\right] = P_{gF}(k,\mu_k)^2 + \left(P_g(k,\mu_k) + n_g^{-1}\right)\left(P_F(k,\mu_k) + P^{1D}(k\mu_k)\; n_{eff}^{-1}\right) \, . \tag{5.2.13}$$

In this approximation, the expected S/N in a bin of (k,μ_k) can be approximated by

$$\left(\frac{S}{N}\right)^2_{Fg} = N_k \, \frac{P_{gF}^2(k,\mu_k)}{P_{gF}(k,\mu_k)^2 + \left(P_g(k,\mu_k) + n_g^{-1}\right)\left(P_F(k,\mu_k) + P^{1D}(k\mu_k)\; n_{eff}^{-1}\right)} \, . \tag{5.2.14}$$

Expected values for the BOSS survey

Here we quantify the previous results for the case of a spectroscopic survey with properties similar as the BOSS survey. As an example of crosscorrelation, we compute the cases of the quasar-Lyα and DLA-Lyα crosscorrelations.

The BOSS survey have an area of $A = 10^4\,\mathrm{deg}^2$, or roughly $A = 5 \times 10^7 (h^{-1}Mpc)^2$, with a depth of $L \sim 10 \times 10^3\, h^{-1}\,\mathrm{Mpc}$ and an effective density of lines of sight $n_{eff} \sim 3 \times 1^{-4} (\,h^{-1}\,\mathrm{Mpc})^{-2}$ (McQuinn and White, 2011).

For the Lyα power spectrum we use the linear bias parameters measured in numerical simulations in McDonald (2003), $b_F = -0.1315$ and $\beta_F = 1.58$, both measured at $z = 2.25$.

We assume quasars have a bias of $b_g = 3$, while DLAs have smaller bias $b_g = 1.5$. We also assume a quasar density of $n_g = 10^{-1} (\,h^{-1}\,\mathrm{Mpc})^{-3}$ and that we detect a DLA in 10% of the lines of sight.

In figure 5.1 we show the S/N estimation for (from top to bottom) the Lyα autocorrelation, quasar-Lyα crosscorrelation, quasar autocorrelation, DLA-Lyα crosscorrelation and DLA autocorrelation, for the values mentioned above. We can see that the S/N of the galaxy-Lyα crosscorrelation is higher than the galaxy autocorrelation, both for the cases of DLAs and quasars.

The linear bias model assumed in this computations breaks down at relatively small values of k and one should restrict to a maximum value k_{max}. One could model a nonlinear correction and calibrate it in numerical simulations in order to push the limit to higher values of k.

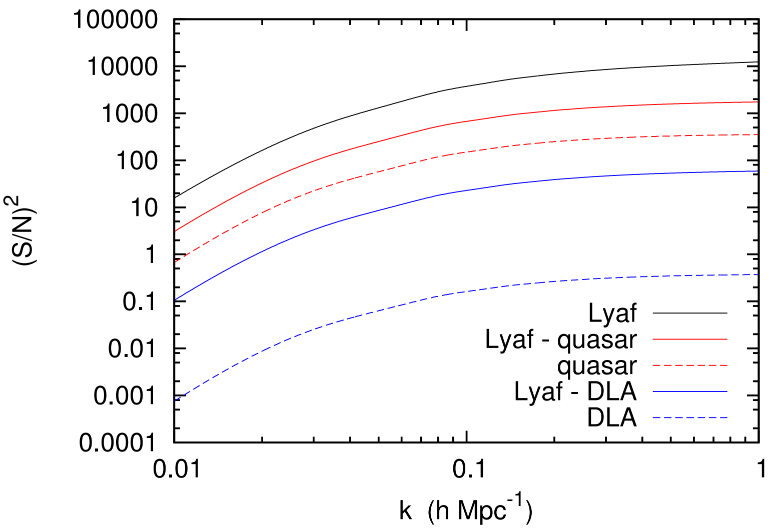

Figure 5.1: Expected S/N for the two examples mentioned in the text, compared to the S/N expected for autocorrelation.

5.2.2 Uncertainty in the Bias Parameters

Here we compute what uncertainties should one expect in the measurement of the bias of objects with their crosscorrelation with the Lyα forest using the Fisher matrix formalism (Tegmark et al., 1997).

The Fisher matrix is defined as

$$F_{ij} = -\left\langle \frac{\partial^2 L}{\partial \lambda_i \partial \lambda_j} \right\rangle = \sum_k N_k \frac{1}{var[P_{Fg}(k,\mu_k)]} \frac{\partial P_{Fg}(k,\mu_k)}{\partial \lambda_i} \frac{\partial P_{Fg}(k,\mu_k)}{\partial \lambda_j} , \tag{5.2.15}$$

where λ_i,λ_j are two of the parameters of the theory.

From the Fisher matrix we can estimate the uncertainty in a given parameter,

$$\sigma_i^2 = (F^{-1})_{ii} . \tag{5.2.16}$$

Example: Quasar bias and $f(\Omega_m)$

Assuming that we know the value of the Lyα forest bias parameters with much higher precision than the parameters we are interested in, we can trivially have an estimate of the uncertainty in the quasar bias, and also in the value of $f(\Omega_m)$ at the redshift. In this case, the only relevant derivatives can be computed analytically,

$$\frac{\partial P_{Fg}}{\partial b_g} = \frac{P_{Fg}}{b_g + f(\Omega_m)\mu_k^2} , \tag{5.2.17}$$

$$\frac{\partial P_{Fg}}{\partial f(\Omega_m)} = \frac{P_{Fg}}{b_g + f(\Omega_m)\mu_k^2}\mu_k^2 \,. \tag{5.2.18}$$

$k_{max}(\,h\,\mathrm{Mpc}^{-1})$	σ_{b_g}	$\sigma_{f(\Omega_m)}$	$\sigma_{b_g}(f(\Omega_m)=1)$
1.0	0.172	0.290	0.089
0.5	0.178	0.304	0.094
0.3	0.189	0.324	0.100
0.1	0.264	0.462	0.143

Table 5.1: Uncertainties in the parameters as a function of k_{max}. In the last column, the uncertainties assuming $f(\Omega_m) = 1$ as predicted by General Relativity if $\Omega_m = 1$

In table 5.1 we show the expected uncertainties using the values mentioned above, for a multiple fit of $b_g - f(\Omega_m)$ and for an individual fit of the quasar bias b_g assuming $f(\Omega_m) \sim 1$ as predicted by General Relativity if $\Omega_m = 1$.

The results are promising and suggest that one should be able to measure the quasar bias with an accuracy better than 10%. If this is the case one could split the data in redshift bins to study the evolution of quasar clustering, or split the data in quasar luminosities to study the mass-luminosity relation of quasars.

5.3 Simple Method to Measure the Crosscorrelation

In order to compute the autocorrelation of any kind of object one needs to know what is its selection function, i.e. the probability to detect an object at a given position. Otherwise one can not trust its estimate of the overdensity δ_g of objects at the point. This is not the case for the Lyα autocorrelation, because the value of δ_F does not depend on the selection function of the background quasar.

When studying the crosscorrelation of the Lyα absorption with any type of object, one might thing that the selection function of the objects is needed. This is indeed the case if one wants to compute the power spectrum or more generally, if one wants to treat the objects as a continuos field instead of a set of discrete points.

Here we show that the crosscorrelation can be computed without any selection function, just computing the averaged value of δ_F at a given separation from an object. We present a simple method to weight the pixels in a quasi-optimal way and present a simple method to compute an approximation for the covariance matrix that is very close to the true covariance in the case of rare objects.

The crosscorrelation of the Lyα absorption field $F(\mathbf{x}) = \bar{F}\,[1+\delta_F(\mathbf{x})]$ with any field of objects (galaxies, quasars, DLAs, etc.) $g(\mathbf{x}) = \bar{g}\,\left[1+\delta_g(\mathbf{x})\right]$.

We define the crosscorrelation function as

$$\xi_{Fg}(\mathbf{r}) = \left\langle \delta_g(\mathbf{x})\,\delta_F(\mathbf{x}+\mathbf{r}) \right\rangle \ , \tag{5.3.1}$$

or equivalently,

$$\langle F(\mathbf{x})\,g(\mathbf{x}+\mathbf{r})\rangle = \bar{F}\,\bar{g}\,\left[1+\xi_{Fg}(\mathbf{r})\right] \ . \tag{5.3.2}$$

Because the galaxy field $g(\mathbf{x})$ can only take the values 0 or 1, the crosscorrelation of this field with any other field will be

$$\langle g(\mathbf{x})\,F(\mathbf{x}+\mathbf{r})\rangle = \bar{g}\ \langle F(\mathbf{r})\rangle_g \ , \tag{5.3.3}$$

and

$$\xi_{gF}(\mathbf{r}) = \left\langle \delta_g(\mathbf{x})\,\delta_F(\mathbf{x}+\mathbf{r}) \right\rangle = \langle \delta_F(\mathbf{r})\rangle_g \ , \tag{5.3.4}$$

where $\langle X(\mathbf{r})\rangle_g$ is the average of any field X over pixels at a distance $\mathbf{r}$ from a galaxy.

5.3.1 Weighted Measurement and Covariances

Here we present a simple method to estimate the crosscorrelation of Lyα absorption with any kind of galaxies, based on the technique described above.

One can estimate the crosscorrelation in a bin $\mathbf{r_A}$ with a weighted average of δ_F in all pixels at a separation from a galaxy that lies inside the bin:

$$\hat{\xi_A} = \frac{\sum_{i\in A} w_i\,\delta_{Fi}}{\sum_{i\in A} w_i} \ . \tag{5.3.5}$$

For instance one can weight the pixels with the inverse of the total variance in each pixel, taking into account both the noise variance and the intrinsic Lyα fluctuations, as suggested by McQuinn and White (2011).

If the bins were infinitely thin (so $\mathbf{r_i} = \mathbf{r_A}$ for all pixels in the bin) it is easy to show that the estimate is unbiased,

$$\left\langle \hat{\xi_A} \right\rangle = \frac{\sum_{i\in A} w_i\ \langle \delta_{Fi}\rangle_g}{\sum_{i\in A} w_i} = \xi_{gF}(\mathbf{r_A}) \ . \tag{5.3.6}$$

The covariance of the measurement in two bins A and B will be

$$\left\langle \hat{\xi_A}\hat{\xi_B} \right\rangle = \frac{\sum_{i\in A}\sum_{j\in B} w_i\,w_j\ \left\langle \delta_{Fi}\,\delta_{Fj}\right\rangle_g}{\sum_{i\in A} w_i\ \sum_{j\in B} w_j} = \frac{\sum_{i\in A}\sum_{j\in B} w_i\,w_j\,C_{ij}}{\sum_{i\in A} w_i\ \sum_{j\in B} w_j} \ , \tag{5.3.7}$$

where C_{ij} is the correlation of pixels i and j, and can be computed from the data itself.

If one wants to model the correlation C_{ij} instead, there are three main terms that should be taken into account: the Lyα autocorrelation $\xi_F(\mathbf{r_{ij}})$, a noise term σ^2_{Ni} for pairs $i = j$ and a contribution from errors in the continuum fitting for pairs of pixels from the same spectrum $\xi_c(r_{ij})$,

$$C_{ij} = \xi_F(\mathbf{r_{ij}}) + \sigma^2_{Ni}\delta^K_{ij} + \xi_c(r_{ij})\delta^D(\mu_{ij}) \, , \tag{5.3.8}$$

where μ_{ij} is the angle cosine relative to the line of sight, and $\delta^D(x)$ is the Dirac delta function.

5.4 Measure DLA Bias from a Lyα Absorption Survey

We apply here the method explained in the previous section to estimate the uncertainty in the bias parameters of DLAs that one could obtain from a quasar survey with similar properties as BOSS.

First we describe the mock spectra used, and then compute the bias parameters for the different realizations.

5.4.1 Mock Spectra

In Chapter 2 we have explained a method to generate mock spectra of Lyα absorption, with any desired one and two-point statistics.

In this study we use the same set of mock spectra as in Chapter 4. These mock spectra were generated with a simplified version of the code that allow a better comparison with the theory. This simplified version do not include redshift evolution, but generate the lines of sight at a fixed redshift of $z_c = 2.6$ instead. The non-linear correction of the Lyα power spectrum used in Chapter 2 is not used here either, allowing us to consider the linear bias approximation as correct.

The set of mocks consist in 50 realizations of a survey with an area of $A = 300\,\mathrm{deg}^2$, and in a redshift range $2.15 < z < 3.5$. Quasars are distributed following the luminosity function of Jiang et al. (2006). We generate spectra for only 75% of the quasars, to obtain a quasar density similar to the observed during the first year of BOSS data $15 - 17\,\mathrm{deg}^{-2}$.

High column density systems have been introduced in the peaks of the optical depth field as described in Chapter 4. The systems were distributed on the 1 % of pixels with higher optical depth, reproducing the observed column density distribution derived by McDonald et al. (2005). Roughly 10% of the spectra contain a Damped Lyα system in the Lyα forest, with a column density higher than $N_{HI} > 10^{20.2}\,\mathrm{cm}^2$. We consider the crosscorrelation of these systems with the Lyα forest. In a real survey the detectability of these systems would depend of the magnitude and redshift of the backlight quasar, but the number of DLAs detected in Slosar et al. (2011) using the first year of BOSS data is also close to 10%.

We simulate the observational noise by adding a gaussian variable to the mock spectra, as described in Chapter 2.

Finally, we rescale the errorbars to simulate the BOSS survey. The total area covered by the survey is roughly 30 times that of our mock survey, with a similar quasar density. Assuming that the variance of the measurement scale with the inverse of the area, we divide the escattering between realizations by a factor $\sqrt{30}$.

5.4.2 Fitting the Bias Parameters

We compute the crosscorrelation in thin bins in r (150 bins of $1\,h^{-1}$ Mpc) and μ (20 bins) using equation 5.3.5. We then compress the information into multipoles, defined as:

$$\xi_{lFg}(r) = \frac{2l+1}{2} \int_{-1}^{1} d\mu \; \xi_F(r,\mu) \; L_l(\mu) \; , \tag{5.4.1}$$

where $L_l(x)$ are the Legendre polynomials and we compute the multipoles for $l = 0, 2, 4$.

In figure 5.2 we show the multipoles measured form the mock spectra. The errorbars have been scaled to mimic the BOSS survey. The expected value has been computed using the Fourier transform of equation 5.2.12. The Lyα bias parameters used are $b_F = -0.177$ and $\beta_F = 1.58$ as in Chapter 4. The value for the linear bias of the DLAs added to the mocks are $b_g = 1.3$, $\beta_g = 1.58$ as suggested also by 4.

Figure 5.2 suggests that one should be able to measure the crosscorrelation of Lyα and DLAs from the BOSS survey, with a significance of $3-\sigma$ up to $r = 40 - 60\,h^{-1}$ Mpc.

Extract bias from the multipoles

In Chapter 4 we describe a method to extract the bias parameters from the multipoles. Here we apply the same formalism to the crosscorrelation.

The redshift space crosscorrelation can be related to the linear matter power spectrum:

$$\xi_{Fg}(r,\mu) = b_F \; b_g \sum_{l=0,2,4} L_l(\mu) \; K_l(\beta_F, \beta_g) \; \xi_l(r) \; , \tag{5.4.2}$$

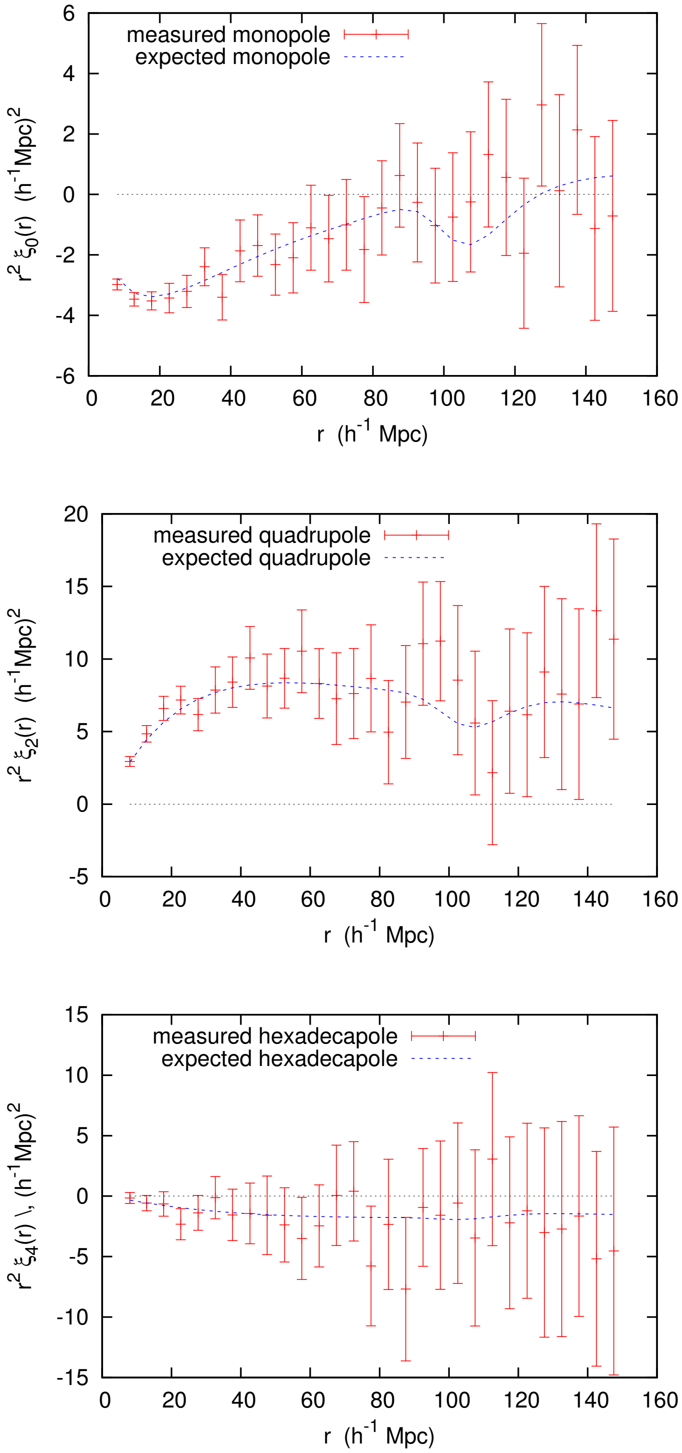

Figure 5.2: Multipoles of the crosscorrelation of DLAs with Lyα forest, from the mock spectra and the expected value. The errorbars are rescaled to mimmic the BOSS survey.

where $\xi_l(r)$ are theoretical functions related to the linear matter power spectrum:

$$\xi_l(r) = \frac{i^{-l}}{2\pi^2} \int_0^\infty dk\, k^2\, j_l(kr)\, P_L(k), \tag{5.4.3}$$

and $j_l(x)$ is the spherical Bessel function of order l. We have also defined the parameters

$$K_0(\beta_F, \beta_g) = 1 + \frac{1}{3}(\beta_F + \beta_g) + \frac{1}{5}\beta_F\, \beta_g\,, \tag{5.4.4}$$

$$K_2(\beta_F, \beta_g) = \frac{2}{3}(\beta_F + \beta_g) + \frac{4}{7}\beta_F\, \beta_g\,, \tag{5.4.5}$$

$$K_4(\beta_F, \beta_g) = \frac{8}{35}\beta_F\, \beta_g\,. \tag{5.4.6}$$

The functions $\xi_l(r)$ are directly related to the observed multipoles of the crosscorrelation in the linear regime:

$$\xi_{lFg}(r) = b_F\, b_g\, K_l(\beta_F, \beta_g)\, \xi_l(r)\,. \tag{5.4.7}$$

Because this relation is scale independent we can compress all the information into a constant for each multipole Chapter 4:

$$X_l \equiv \frac{\int_{r_1}^{r_2} dr\, \xi_{lFg}(r)\, w(r)}{\int_{r_1}^{r_2} dr\, \xi_l(r)\, w(r)} = b_F\, b_g\, K_l(\beta_F, \beta_g)\,, \tag{5.4.8}$$

where we weight every r bin with the inverse of the dispersion between the value measured in each realization $w(r)$. Since the linear bias approximation breaks down on small scales, we restrict our analysis to $r > r_1 = 10\, h^{-1}$ Mpc. We use a value of $r_2 = 80\, h^{-1}$ Mpc, but the results are not sensitive to this choice.

Using several mock realizations of the survey one can compute the mean values $\hat{X}_i = \langle X_i \rangle$ and the elements of the covariance matrix of C_{ij},

$$C_{ij} = \langle X_i X_j \rangle - \hat{X}_i \hat{X}_j\,, \tag{5.4.9}$$

and fit the best value of the DLA bias parameters b_g, β_g for the true values of the Lyα bias parameters b_F, β_F.

$$\chi^2(b_g, \beta_g) = (\hat{X}_i - b_F\, b_g\, K_i(\beta_F, \beta_g))\, C_{ij}^{-1}\, (\hat{X}_j - b_F\, b_g\, K_j(\beta_F, \beta_g))\,. \tag{5.4.10}$$

We plot in figure 5.3 the result of the fit. Despite of the high degeneration between both parameters, we are able to constrain the value of b_g in $1 < b_g < 1.6$ ($1-\sigma$ uncertainties).

In figure 5.4 we redo the same analysis but now with mocks where the DLAs have been added only in the 0.2 % of pixels with higher optical depth, producing DLAs with a larger bias. We can see that the uncertainties are similar to the previous case, even though now the signal is larger.

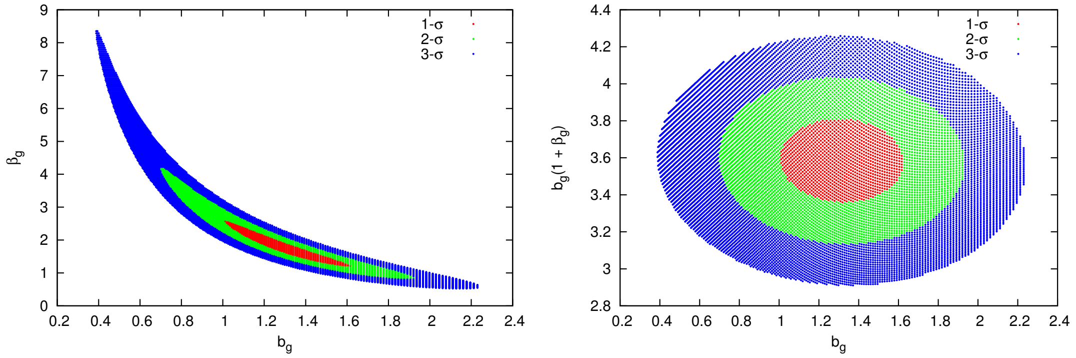

Figure 5.3: $1-\sigma, 2-\sigma, 3-\sigma$ contours in the b_g , β_g plane (left) and in the b_g, $b_g(1+\beta_g)$ plane.

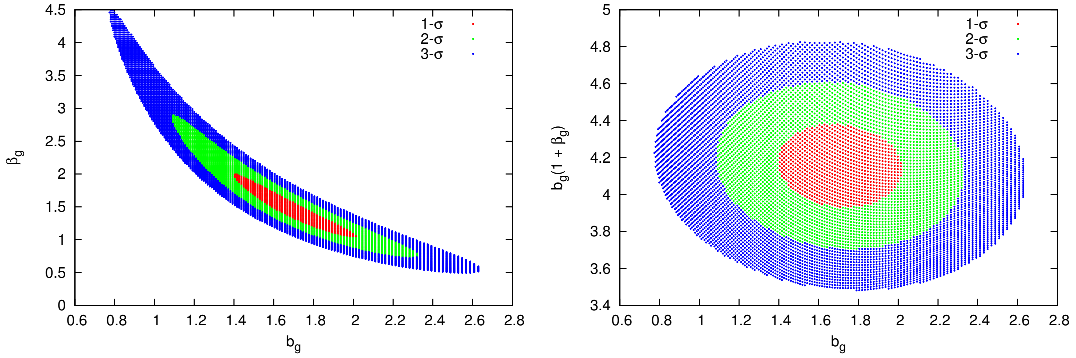

Figure 5.4: Same figure than above, but with more clustered DLAs.

5.4.3 Mass of the Host Halo

Given a mass function one can have a relation between the mass of a halo and its linear bias. One could use the measure of the bias parameter to constraint the mass of the host halos (Sheth and Tormen, 1999),

$$b_h(M,z) = 1 + \frac{a\,\nu^2(M,z) - 1}{\delta_c} + \frac{2p}{\delta_c\,(1 + (a\,\nu^2(M,z))^p)}\,, \tag{5.4.11}$$

where we have defined

$$\nu(M,z) = \frac{\delta_c}{\sigma(M,z)} = \frac{\delta_c}{\sigma(M,0)}\frac{G(0)}{G(z)}\,, \tag{5.4.12}$$

and $G(z)$ is the linear growth factor and $\sigma(M)^2$ is the variance of linear fluctuations at scales $M = \rho_m \frac{4\pi}{3} R^3$.

In figure 5.5 we plot the relation for $a = 1$, $p = 0$ (Press & Shechter relation) and for the $a = 0.75$, $p = 0.3$ (Sheth & Tormen relation). The blue lines show the uncertainties measured from the mocks if $b_g = 1.3$, while the red lines show the same uncertainty

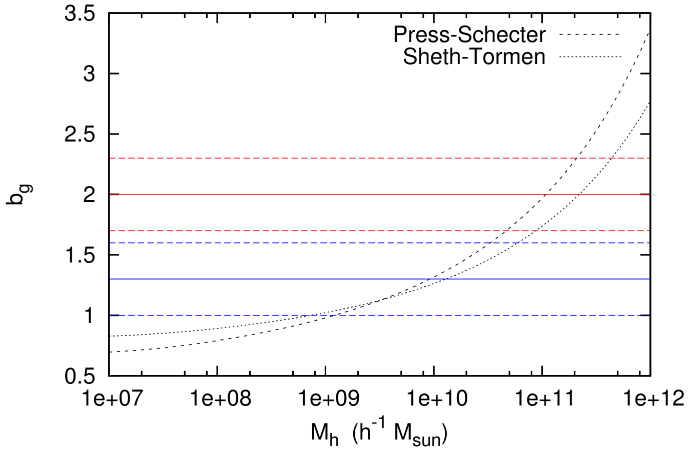

Figure 5.5: Constraints in the host halo mass from the bias measurement.

but for a higher value of $b_g = 2.0$. Because the mass - bias relation flattens for small masses, if the bias of DLAs is small one would only be able to constraint an upper limit to the mass of the host halos. On the other hand, if the bias is high one should be able to measure the order of magnitude for the mass of the halos hosting DLAs.

5.5 Conclusions

We have shown that the crosscorrelation with the Lyα absorption is a very powerful statistics to study the clustering of rare objects. Using simple Fisher matrix estimation we predict that the average bias of quasars could be computed to an accuracy better than 10%.

We have presented a simple method to compute the crosscorrelation without the need of the selection function of the objects, by computed the averaged value of the transmitted fraction around the objects. The covariance of the different bins measured could also be computed from the data itself.

We compute the expected uncertainty in the measurement of the bias of DLAs by using the previous method in a set of mock spectra. The results show that using the BOSS survey one could measure the bias parameters of DLAs with an accuracy of $\sim$ 10%.

There are several caveats that have not been studied here, and should been taken into account when analyzing the data:

- The crosscorrelation of quasars with Lyα would be affected by the errors in the measure of the quasar redshift. The correlation on scales similar and smaller than the typical redshift error would be highly smoothed. A typical value for the error in BOSS quasars is $\sim 10\,h^{-1}$ Mpc.

- Since the detectability of DLAs depend on the amount of Lyα absorption, the value of β_g and the shot noise term in the variance of the power spectrum may not hold.

- The value of k_{max} that limits the validity of the linear bias approximation should be calibrated in numerical simulations. In the case of the crosscorrelation of quasar and the Lyα forest, ionization effects should also been taken into account.

- In this analysis we have assumed that we know with high accuracy the value of the Lyα bias parameters. This should be the case after the whole BOSS survey is completed, but a joint analysis could be necessary.

Chapter 6

Conclusions and Future Perspectives

6.1 Conclusions

Our hopes to detect the Baryon Acoustic Oscillations (BAO) signature in the correlation of the Lyα forest have been placed on solid ground with the first measurement of the Lyα correlation function on cosmologically large scales by the Baryon Oscillations Spectroscopic Survey (BOSS, Slosar et al., 2011). In this paper, we show that the redshift distorted correlation function can be explained with linear bias theory, and we measure the values of b_δ and β.

Simulated data sets are crucial for any analysis and interpretation of the data. The method developed in Chapter 2 allows us to generate multiple realizations of the survey and to introduce multiple layers of complexity in order to obtain spectra as realistic as possible. I show that to optimize the BOSS survey is better to increase the survey area than the exposure time, in agreement with previous studies.

Simulated spectra generated with the code developed in Chapter 2 played a crucial role in the data analysis of Slosar et al. (2011). In Chapter 3 I summarize the publication and highlight my personal contribution and explain how mock catalogs were used to test the data analysis code and the algorithm to compute the covariance matrix.

A study of the effect of high column density systems (HCDs) in the correlation function of Lyα absorption is presented in Chapter 4. I compute the systematic effect in the measure of the Lyα bias parameters and propose an explanation for the low value of β measured in Slosar et al. (2011). We test the results by introducing HCD systems into mock spectra and study the changes in the correlation function.

Finally, I show that the crosscorrelations of the Lyα forest with galaxies, quasars or different absorption systems will be measurable in the BOSS survey. These measurements will provide information on the formation mechanisms and evolution of these systems. For instance, the bias of the halos hosting Damped Lyα systems (DLAs) could be measured with a $\sim 30\%$ accuracy.

6.2 Future Perspectives

The BOSS survey is meant to be a turning point in the study of the Lyα forest. The large number of quasar spectra that are being obtained will allow us to study a long list of exciting projects. The results obtained from these studies will confirm the Lyα forest as one of the most promising techniques to study not only the large scale structure, but also galaxy formation and the evolution of the IGM.

The main goal of the BOSS survey is to detect the BAO feature in the correlations of the Lyα forest. The forecasts predict an accuracy of few percent in the Hubble parameter $H(z)$ and in the angular diameter distance $d_A(z)$ at a redshift of $z = 2 - 3$. Combining the results with CMB data and the detection of BAO at low redshift using galaxies will provide important constraints on the cosmological parameters, especially on the curvature of the Universe.

Using the full shape of the correlations (either in Fourier or real space) and the one-dimensional power spectrum, one can obtain even stronger constraints. A new upper limit on the neutrino masses will be obtained, but its difficult to quantify it because it will be limited by systematics in the measure.

The crosscorrelations of the Lyα forest with different collapsed systems will provide new information on the formation mechanisms of the systems. The crosscorrelation with quasars on small scales can be used to obtain information on the lifetime and isotropy of its radiation. On large scales we will be able to measure its linear bias parameters and study its dependence on luminosity, redshift, colour, variability, etc. The total number of DLAs identified in BOSS spectra could be larger than 10000. This would allow us to obtain a measure of the crosscorrelation with the Lyα forest, and measure an estimation of the bias parameter of their host halos.

APPENDIX A

Bias of the Gaussian Peaks

A.1 Biases of the Gaussian Field

Given any transformation $F(\delta_g)$ we can compute the relation between the correlation function of the Gaussian field ξ_g and the correlation function of the flux field ξ_F [1] :

$$\begin{aligned}\bar{F}^2(1+\xi_F(r_{12})) &= \langle F_1 F_2 \rangle \qquad \text{(A.1.1)}\\ &= \int_0^1 dF_1 \int_0^1 dF_2 \; p_F(F_1, F_2) \; F_1 \; F_2 \\ &= \int_{-\infty}^{\infty} d\delta_{g1} \int_{-\infty}^{\infty} d\delta_{g2} \; p_g(\delta_{g1}, \delta_{g2}) \; F_1 \; F_2 \\ &= \int_{-\infty}^{\infty} d\delta_{g1} \int_{-\infty}^{\infty} d\delta_{g2} \frac{e^{-\dfrac{\delta_{g1}^2+\delta_{g2}^2-2\delta_{g1}\delta_{g2}\xi_g(r_{12})}{2(1-\xi_g^2(r_{12}))}}}{2\pi\sqrt{1-\xi_g^2(r_{12})}} \; F(\delta_{g1}) \; F(\delta_{g2}) \; .\end{aligned}$$

The Gaussian variables $\delta_{g1} = \delta_g(x_1)$ and $\delta_{g2} = \delta_g(x_2)$ are normal variables ($\langle \delta_{gi} \rangle = 0$, $\left\langle \delta_{gi}^2 \right\rangle = 1$) and are correlated by $\xi_g(r_{12})$. We can always redefine these variables as a linear combination of 2 independents normal variables y_1,y_2:

$$\delta_{g1} = y_1 \qquad \delta_{g2} = \xi_g \; y_1 + \sqrt{1-\xi_g^2} \; y_2 \; . \qquad \text{(A.1.2)}$$

These variables satisfy now:

$$\langle y_i \rangle = 0, \qquad \langle y_i y_j \rangle = \delta_{ij}^K, \qquad p_g(y_1, y_2) = p_g(y_1) \; p_g(y_2) \; . \qquad \text{(A.1.3)}$$

The expression for $\xi_F(\mathbf{r_{12}})$ is now:

$$\bar{F}^2(1+\xi_F(\mathbf{r_{12}})) = \int_{-\infty}^{\infty} dy_1 \; p_g(y_1) \; F(y_1) \int_{-\infty}^{\infty} dy_2 \; p_g(y_2) \; F(\delta_{g2}(y_1, y_2, \xi_g(r_{\mathbf{r_{12}}}))) \; . \quad \text{(A.1.4)}$$

[1] here we do not distinguish between F and F_α because no HCD system is involved in the computation

In the limit $\xi_g \ll 1$, we can Taylor expand the expression for $F(\delta_{g2})$ around y_2, i.e. around $\xi_g = 0$:

$$\begin{aligned} F(\delta_{g2}(y_1, y_2, \xi_g)) &= F(\xi_g\, y_1 + \sqrt{1-\xi_g^2}\, y_2) \\ &= F(y_2) + \frac{dF}{d\xi_g}\, \xi_g \\ &= F(y_2) + \frac{dF}{d\tau}\frac{d\tau}{d\delta_{g2}}\frac{d\delta_{g2}}{d\xi_g}\, \xi_g \\ &= F(y_2) - F(y_2)\frac{d\tau}{d\delta_{g2}} y_1\, \xi_g \\ &= F(y_2)\left(1 - \frac{d\tau}{d\delta_{g2}} y_1\, \xi_g\right) , \end{aligned} \tag{A.1.5}$$

and the equation that relates both correlation functions is now:

$$\begin{aligned} \bar{F}^2(1+\xi_F(\mathbf{r_{12}})) &= \langle F_1 F_2 \rangle \\ &= \int_{-\infty}^{\infty} dy_1\; p_g(y_1)\; F(y_1) \int_{-\infty}^{\infty} dy_2\; p_g(y_2)\; F(\delta_{g2}(y_1, y_2, \xi_g(\mathbf{r_{12}})) \\ &= \int_{-\infty}^{\infty} dy_1\; p_g(y_1)\; F(y_1) \int_{-\infty}^{\infty} dy_2\; p_g(y_2)\; F(y_2)\left(1 - \frac{d\tau}{dy_2}\, y_1\, \xi_g\right) \\ &=\bar{F}^2\left(1 + \left(\frac{b_F}{b_g}\right)^2 \xi_g\right) , \end{aligned} \tag{A.1.6}$$

where $\tau = -ln(F)$ is the optical depth and b_F and b_g are the bias parameters for the variable δ_F and δ_g respectively:

$$\left(\frac{b_F}{b_g}\right)^2 = \frac{1}{\bar{F}^2}\int_{-\infty}^{\infty} dy_1\; p_g(y_1)\; F(y_1)\; y_1 \int_{-\infty}^{\infty} dy_2\; p_g(y_2)\; F(y_2)\; \frac{d\tau}{dy_2} . \tag{A.1.7}$$

We have shown that on large scales $\xi_g \ll 1$ both correlations are just proportional and share the same redshift distortion parameter β.

A.2 Bias of the Peaks

In section 4.4 we describe a method to distribute HCD systems in the peaks of Lyα absorption by choosing only the pixels with an optical depth above a threshold τ_c (or equivalently, above a threshold δ_{gc} in the Gaussian variable used to generate the optical depth). This threshold sets what fraction ν of pixels are candidate to host a DLA:

$$\nu = \int_{\delta_{gc}}^{\infty} d\delta_g\; p_g(\delta_g) , \tag{A.2.1}$$

where $p_g(\delta_g)$ is the Gaussian probability distribution.

The correlation function of these peaks ξ_h is related to the probability of having a peak both at $\mathbf{x_1}$ and at $\mathbf{x_2}$:

$$\begin{aligned} p(\delta_{g1} > \delta_{gc},\, \delta_{g2} > \delta_{gc}) &= \nu^2\,(1 + \xi_h(\mathbf{r_{12}})) \\ &= \int_{\delta_{gc}}^{\infty} d\delta_{g1} \int_{\delta_{gc}}^{\infty} d\delta_{g2}\; p_g(\delta_{g1}, \delta_{g2}) \\ &= \int_{\delta_{gc}}^{\infty} d\delta_{g1} \int_{\delta_{gc}}^{\infty} d\delta_{g2}\; \frac{e^{-\dfrac{\delta_{g1}^2 + \delta_{g2}^2 - 2\delta_{g1}\delta_{g2}\xi_g(\mathbf{r_{12}})}{2(1-\xi_g^2(\mathbf{r_{12}}))}}}{2\pi\sqrt{1-\xi_g^2(\mathbf{r_{12}})}} \,. \end{aligned} \tag{A.2.2}$$

We can now use express δ_{g1},δ_{g2} as a function of the same independent normal variables y_1, y_2 defined in section A.1:

$$\delta_{g1} = y_1 \qquad \delta_{g2} = \xi_g\, y_1 + \sqrt{1-\xi_g^2}\, y_2\,, \tag{A.2.3}$$

to obtain the expression

$$\begin{aligned} \nu^2\,(1+\xi_h) &= \int_{\delta_{gc}}^{\infty} d\delta_{g1} \int_{\delta_{gc}}^{\infty} d\delta_{g2}\; p_g(\delta_{g1}, \delta_{g2}) \\ &= \int_{\delta_{gc}}^{\infty} dy_1\; p_g(y_1) \int_{\frac{\delta_{gc}-y_1\xi_g}{\sqrt{1-\xi_g^2}}}^{\infty} dy_2\; p_g(y_2) \\ &= \int_{\delta_{gc}}^{\infty} dy_1\; p_g(y_1) \left(\int_{\frac{\delta_{gc}-y_1\xi_g}{\sqrt{1-\xi_g^2}}}^{\delta_{gc}} dy_2\; p_g(y_2) + \int_{\delta_{gc}}^{\infty} dy_2\; p(y_2) \right) \\ &\sim \int_{\delta_{gc}}^{\infty} dy_1\; p_g(y_1)\, \left((\delta_{gc} - (\delta_{gc} - y_1\,\xi_g))\; p_g(\delta_{gc}) + \nu \right) \\ &\sim \nu^2 \left(1 + \left(\frac{b_h}{b_g}\right)^2 \xi_g \right) \end{aligned} \tag{A.2.4}$$

where b_g and b_h are the bias parameters of the Gaussian field and the peaks respectively:

$$\left(\frac{b_h}{b_g}\right)^2 = \frac{p_g(\delta_{gc})}{\nu^2} \int_{\delta_{gc}}^{\infty} dy_1\; p_g(y_1)\, y_1\,. \tag{A.2.5}$$

We have shown that the correlation of the peaks is proportional to the correlation of the Gaussian field and share the same redshift distortion parameter β.

APPENDIX B

Slosar et al. 2011

Prepared for submission to JCAP

The Lyman-α forest in three dimensions: measurements of large scale flux correlations from BOSS 1st-year data

Anže Slosar,[a] Andreu Font-Ribera,[b] Matthew M. Pieri,[c,d] James Rich,[e] Jean-Marc Le Goff,[e] Éric Aubourg,[f,e] Jon Brinkmann,[g] Nicolas Busca,[f] Bill Carithers,[h] Romain Charlassier,[e] Marina Cortês,[h] Rupert Croft,[i] Kyle S. Dawson,[j] Daniel Eisenstein,[k] Jean-Christophe Hamilton,[f] Shirley Ho,[h] Khee-Gan Lee,[l] Robert Lupton,[l] Patrick McDonald,[h,a] Bumbarija Medolin,[m] Jordi Miralda-Escudé,[n,o] Adam D. Myers,[p,q] Robert C. Nichol,[r] Nathalie Palanque-Delabrouille,[e] Isabelle Pâris,[s] Patrick Petitjean,[s] Yodovina Piškur,[l] Emmanuel Rollinde,[s] Nicholas P. Ross,[h] David J. Schlegel,[h] Donald P. Schneider,[t] Erin Sheldon,[a] Benjamin A. Weaver,[u] David H. Weinberg,[d] Christophe Yeche,[e] Donald G. York[v,w]

[a]Brookhaven National Laboratory, Blgd 510, Upton NY 11375, USA

[b]Institut de Ciències de l'Espai (CSIC-IEEC), Campus UAB, Fac. Ciències, torre C5 parell 2, Bellaterra, Catalonia

[c]Center for Astrophysics and Space Astronomy, University of Colorado, 389 UCB, Boulder, Colorado 80309, USA

[d]Department of Astronomy, Ohio State University, 140 West 18th Avenue, Columbus, OH 43210, USA

[e]CEA, Centre de Saclay, IRFU, 91191 Gif-sur-Yvette, France

[f]APC, Université Paris Diderot-Paris 7, CNRS/IN2P3, CEA, Observatoire de Paris, 10, rue A. Domon & L. Duquet, Paris, France

[g]Apache Point Observatory, P.O. Box 59, Sunspot, NM 88349,USA

[h]Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA.

[i]Bruce and Astrid McWilliams Center for Cosmology, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[j]University of Utah, Dept. of Physics & Astronomy, 115 S 1400 E, Salt Lake City, UT 84112, USA

[k]Harvard College Observatory, 60 Garden St., Cambridge MA 02138, USA

[l]Department of Astrophysical Sciences, Princeton University, Princeton, New Jersey 08544, USA

[m]104-20 Queens Blvd #17A, Forest Hills, NY 11375, USA

[n]Institució Catalana de Recerca i Estudis Avançats, Barcelona, Catalonia

[o]Institut de Ciències del Cosmos, Universitat de Barcelona/IEEC, Barcelona 08028, Catalonia

[p]Department of Astronomy, MC-221, University of Illinois, 1002 West Green Street, Urbana, IL 61801, USA

[q]Department of Physics and Astronomy, University of Wyoming, Laramie, WY 82071, USA

[r]Institute of Cosmology and Gravitation (ICG), Dennis Sciama Building, Burnaby Road, Univ. of Portsmouth, Portsmouth, PO1 3FX, UK.

[s]Université Paris 6 et CNRS, Institut d'Astrophysique de Paris, 98bis blvd. Arago, 75014 Paris, France

[t]Department of Astronomy and Astrophysics, The Pennsylvania State University, 525 Davey Lab, University Park, PA 16802, USA

[u]Center for Cosmology and Particle Physics, New York University, New York, NY 10003 USA

[v]Department of Astronomy and Astrophysics, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, USA

[w]Enrico Fermi Institute, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, USA

E-mail: anze@bnl.gov

Abstract. Using a sample of approximately 14,000 $z > 2.1$ quasars observed in the first year of the Baryon Oscillation Spectroscopic Survey (BOSS), we measure the three-dimensional correlation function of absorption in the Lyman-α forest. The angle-averaged correlation function of transmitted flux ($F = e^{-\tau}$) is securely detected out to comoving separations of 60 h^{-1} Mpc, the first detection of flux correlations across widely separated sightlines. A quadrupole distortion of the redshift-space correlation function by peculiar velocities, the signature of the gravitational instability origin of structure in the Lyman-α forest, is also detected at high significance. We obtain a good fit to the data assuming linear theory redshift-space distortion and linear bias of the transmitted flux, relative to the matter fluctuations of a standard ΛCDM cosmological model (inflationary cold dark matter with a cosmological constant). At 95% confidence, we find a linear bias parameter $0.16 < b < 0.24$ and redshift-distortion parameter $0.44 < \beta < 1.20$, at central redshift $z = 2.25$, with a well constrained combination $b\,(1+\beta) = 0.336\pm0.012$. The errors on β are asymmetric, with $\beta = 0$ excluded at over 5σ confidence level. The value of β is somewhat low compared to theoretical predictions, and our tests on synthetic data suggest that it is depressed (relative to expectations for the Lyman-α forest alone) by the presence of high column density systems and metal line absorption. These results set the stage for cosmological parameter determinations from three-dimensional structure in the Lyman-α forest, including anticipated constraints on dark energy from baryon acoustic oscillations.

Keywords: cosmology, Lyman-α forest, large scale structure, dark energy

Contents

1 Introduction

Early spectra of high-redshift ($z > 2$) quasars showed ubiquitous absorption lines blueward of their Lyman-α emission, which were identified as arising primarily from Lyman-α absorption by intervening concentrations of neutral hydrogen [1]. While early models described the absorbers as discrete clouds analogous to those in the interstellar medium, a combination of theoretical and observational advances in the mid-1990s led to a revised view of the forest as a continuous phenomenon, analogous to Gunn-Peterson [2] absorption but tracing an inhomogeneous, fluctuating intergalactic medium (see, e.g., the review by [3]). This revision of theoretical understanding also transformed the promise of the Lyman-α forest as a tool for cosmology, by showing that the forest traces the distribution of dark matter in the high-redshift universe in a relatively simple way. Adopting this "continuous medium" view, several groups [4–9] measured the power spectrum of Lyman-α forest flux in successively larger samples of quasar spectra and used it to constrain the power spectrum of the underlying dark matter distribution. The most recent of these measurements [9], using a large quasar sample from the Sloan Digital Sky Survey (SDSS; [10–17]), have provided some of the strongest constraints on inflation, neutrino masses, and the "coldness" of dark matter, especially when combined with data sets that probe other redshifts and/or physical scales (e.g., [18–23]). However, while the underlying density field is three-dimensional, all of these cosmological analyses have treated the forest as a collection of independent 1-dimensional maps.

This paper presents measurements of Lyman-α forest flux correlations across parallel lines of sight with the widest separation reached so far, taking advantage of the large sample of high-redshift quasars observed during the first year of BOSS, the Baryon Oscillation Spectroscopic Survey of SDSS-III [24].

Measurements of correlated absorption towards gravitational lenses and closely separated quasar pairs [25–28] provided some of the critical observational evidence for the revised understanding of the Lyman-α forest. These transverse correlation measurements implied a coherence scale of several hundred h^{-1} kpc (comoving) for individual absorbing structures. The large sizes and low neutral column densities in turn implied that the absorbing gas has low densities and is highly photoionized by the intergalactic UV background, making the total baryon content of the forest a large fraction of the baryons allowed by Big Bang nucleosynthesis [29]. Hydrodynamic cosmological simulations naturally explained these large coherence scales and many other statistical properties of the forest [30–33], with typical absorption features arising in filamentary structures of moderate overdensity, $\delta \equiv \rho/\bar{\rho} - 1 \sim 20$. Detailed investigation of these simulations revealed an approximate description of the forest that is both surprisingly simple and surprisingly accurate [34, 35]: the Lyman-α forest arises in gas that traces the underlying dark matter distribution, except for small scale pressure support, with the neutral hydrogen fraction determined by photoionization equilibrium, and with a power-law temperature-density relation $T \propto (\rho/\bar{\rho})^{\gamma-1}$. This relation is established by the competition of photoionization heating and adiabatic cooling [36]. In this approximation, the transmitted flux fraction F is related to the dark matter overdensity δ by

$$F = e^{-\tau} = \exp\left[-A(1+\delta)^{2-0.7(\gamma-1)}\right] , \tag{1.1}$$

where the temperature-density slope $(\gamma - 1)$ depends on the inter-galactic medium (IGM) reionization history, and the constant A depends on redshift and on a variety of physical parameters (see [37] for a recent discussion). On large scales, the three-dimensional power spectrum of the field $\delta_F \equiv F/\bar{F} - 1$ should have the same shape as the power spectrum of $\delta = \rho/\bar{\rho} - 1$ [4, 5, 35]. Thermal motions of atoms smooth spectra on small scales, producing a turnover in the one-dimensional flux power spectrum at high k. (In practice, peculiar velocities, which we discuss next, produce a turnover at larger scales than thermal motions.)

The most significant correction to equation (1.1) is from peculiar velocities, which shift the apparent locations of the absorbing neutral hydrogen in the radial direction. On large scales, the power spectrum should approximately follow the linear theory model of redshift-space distortions,

$$P_F(k, \mu_k) = b^2 P_L(k)(1 + \beta\mu_k^2)^2 , \tag{1.2}$$

where $P_L(k)$ is the real-space linear power spectrum and μ_k is the cosine of the angle between the wavevector $\mathbf{k}$ and the line of sight. The bias factor b of the forest is the bias factor of the contrast of the flux fluctuations and not the bias factor of the neutral hydrogen. It is typically low because the full range of density variations is mapped into the transmitted flux range $0 < F < 1$ (b is actually technically negative, because overdensities in mass produce smaller F, but this is irrelevant to our paper so we will just quote $|b|$). While the functional form in Eq. 1.2 is that of Kaiser [38], the parameter β does not have the same interpretation, coming from a more general linear theory calculation of redshift-space distortions in the case where the directly distorted field, in this case optical depth, undergoes a subsequent non-linear transformation, in this case $F = \exp(-\tau)$ [5, 39]. For a galaxy survey, $\beta \approx [\Omega_m(z)]^{0.55}/b$, but for the Lyman-$\alpha$ forest β can vary independently of Ω_m and b, i.e., it is generally a free parameter.

The simulations of [39] suggest an approximate value of $\beta \approx 1.47$ at $z \approx 2.25$ (interpolating over the parameter dependences in Table 1 of [39], since the central model there is antiquated). Lower resolution calculations of [40] ($500h^{-1}$kpc mean particle spacing and particle-mesh (PM) cell size) found a lower value $\beta \sim 1$. Unpublished work by Martin White demonstrates that the value of β predicted by PM simulations with different smoothing applied to the mass density field decreases with increasing smoothing length (based simulations with 73 h^{-1}kpc resolution, and in agreement with the $\sim$ 180 h^{-1}kpc resolution simulation in [41]). Reference [39] concurred that low resolution simulations produce lower β, and agree with White on the value at similar smoothing, so the fundamental outstanding question for predicting β is apparently how smooth is the gas in the IGM ([39] quantifies other parameter dependences, but none of them make so much difference, given the relatively small uncertainty in those parameters, for vanilla models at least). The smoothing scale of the IGM is determined mostly by the pressure, i.e., Jeans smoothing, which is determined by the temperature history of the gas [42, 43]. A linear theory calculation of the smoothing scale for reasonable thermal history, following [42], predicts $\sim 100\,h^{-1}$ kpc or slightly more (this is the rms smoothing length for a Gaussian kernel applied to the density field). [39] used the hydro-PM (HPM) approximation of [42] to include pressure in the non-linear evolution of otherwise PM simulations and found results consistent with smoothing PM by a smaller amount, $\sim 40\,h^{-1}$ kpc. We know of no reason to doubt the qualitative accuracy of the HPM simulations, but ultimately fully hydrodynamic simulations will be required to decisively compute the expected value of β for a given model. [44] plotted the 3D power spectrum from a hydro simulation, and suggested β could be ~ 1, but their single $25\,h^{-1}$ Mpc simulation box had too few modes in the comfortably linear regime to assess the accuracy of this result. Finally, even if we were sure of our calculation for a given thermal history, there is some uncertainty due to uncertainty about the thermal history, especially the redshift of reionization (we have not quantified how large this uncertainty is).

At the percent level needed to interpret 1D power spectrum measurements, the value of the bias parameter b is well known to depend on the mean absorption level (shown directly in [39]), which is difficult to determine accurately; however, at the level of this paper it is actually quite precisely predicted, with [39] and [41] (verified by Martin White, private communication) agreeing on a value $\sim$ 0.13 to $\sim$ 10% with less sensitivity to smoothing than for β ([40] did find a higher value $\sim$ 0.18 for their much lower $500h^{-1}$kpc resolution simulation).

In addition to quasar pair measurements, there have been some studies of closely spaced groupings of quasars that provide hints of three-dimensional structure in the Lyman-α forest

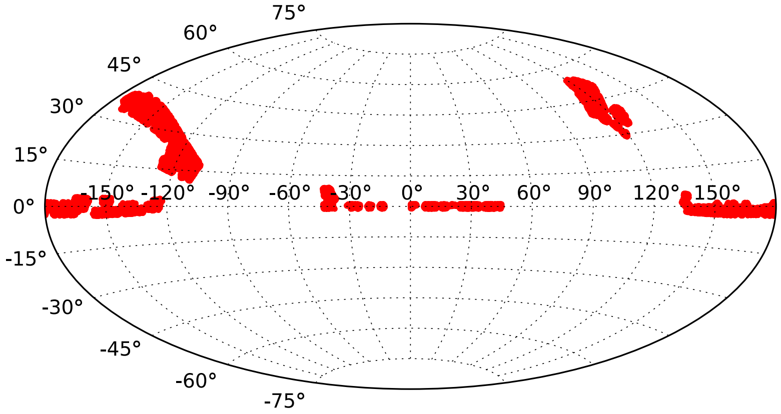

Figure 1. Survey area of the quasars used in this paper in equatorial coordinates, in the Aitoff projection.

[45–53]. However, large scale flux correlations are weak, so detecting them with high significance across widely separated sightlines requires a large and dense quasar sample. By design, the BOSS quasar survey provides precisely such a sample, probing a large comoving volume with ~ 15 sightlines per deg^2.

The ultimate goal of the BOSS survey is to measure the baryon acoustic oscillation (BAO) feature to high precision in the *galaxy* redshift survey at $z < 0.7$ and Lyman-α forest at $z \approx 2.5$ [24] (the possibility of measuring the BAO feature using the Lyman-α forest was suggested in [39] and some preliminary calculations were done by [54], but the potential of the measurement was not really quantified until [55] (see also [40, 41, 56]). These measurements will be used as a standard ruler to measure the angular diameter distance $D_A(z)$ and Hubble parameter $H(z)$. The first proof-of-concept paper for the galaxy part of the survey is [57], and this paper attempts to do the same for the Lyman-α forest part of the survey. The sample of 14,000 quasars analyzed in this paper is too small to yield a confident detection of the BAO feature, but it does allow the first precise measurements of three-dimensional structure in the Lyman-α forest. The basic statistic that we use is the *flux correlation function*

$$\xi_F(r, \mu) = \langle \delta_F(\mathbf{x}) \delta_F(\mathbf{x} + \mathbf{r}) \rangle \ , \tag{1.3}$$

which is the Fourier transform of the power spectrum (1.2). Our measurements provide novel tests of the basic cosmological understanding of high-redshift structure and the nature of the Lyman-α forest.

The next section describes our data sample, based on quasars observed by BOSS between December 2009 and July 2010. During these first few months of BOSS, which included commissioning of new hardware, software, and observing procedures, data taking was relatively inefficient. Our sample of 14,598 $z > 2.1$ quasars over ~ 880 square degrees is less than 10% of the anticipated final sample of 150,000 quasars selected over 10,000 deg^2, but it is already comparable to the total number of $z > 2.1$ quasars found by SDSS-I and II [58]. Section 3 describes our methods for creating realistic synthetic data sets, which are crucial to testing our measurement and error estimation procedures and which we use as a source of

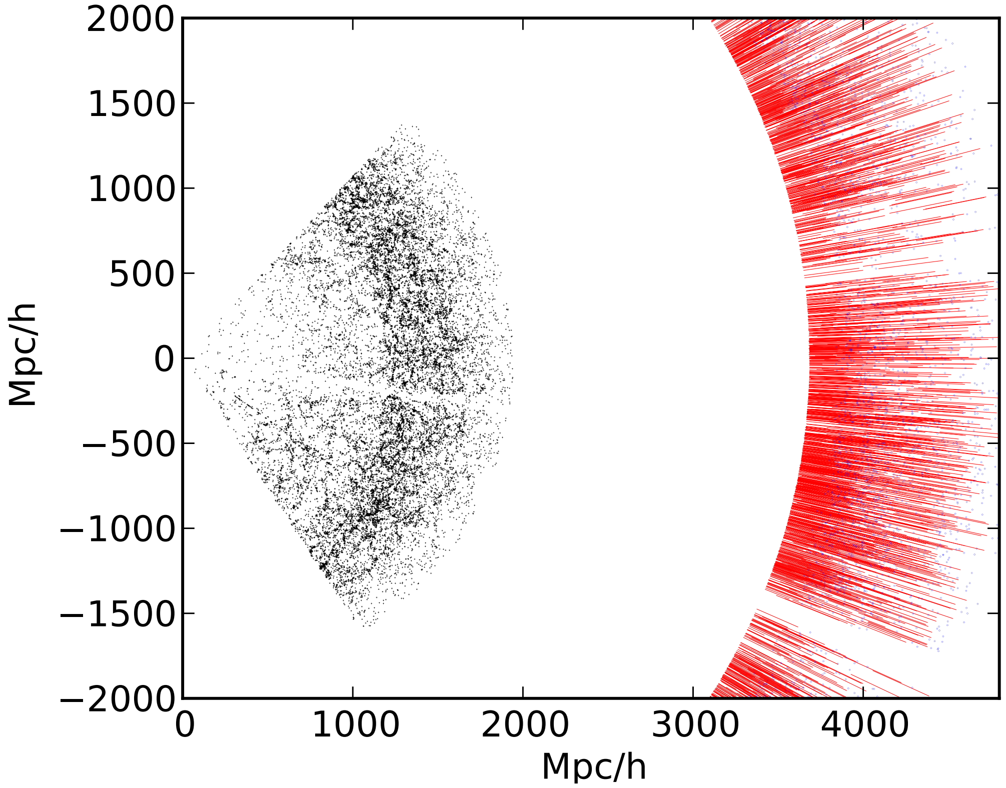

Figure 2. The geometry of the BOSS survey for a thin slice in the equatorial plane. Our galaxy is at the origin. The dark dots are the galaxies measured in the BOSS survey and the blue markers show the positions of quasars whose Lyman-α forests we use. The actual Lyman-α forest regions are shown in as red lines. Apparent differences between geometry of quasar and galaxy distribution arise from small differences in slice thickness and time-span.

theoretical predictions for comparison to our measurements. Section 4 describes our analysis procedures including removal of continuum, estimation of the flux correlation function and its statistical errors, and model fitting, presenting detailed tests of these procedures on the synthetic data. Section 5 presents the main results of the BOSS analysis. We describe several internal tests for systematic errors in §6, and we summarize our results in §7 with final remarks given §8.

2 Data sample

In this section we give a brief overview of the BOSS Quasar data sample that was used in our analysis, and the French Participation Group (FPG) visual inspections of the BOSS spectra. The quasar target selection for the first year of BOSS observations is described in detail in [59], which draws together the various methods presented in [60, 61] and [62]. In summary, selecting $z > 2.1$, and in particular $2.1 < z < 3.5$ objects, has always been challenging due to the colours of quasars at these redshifts crossing over the stellar locus in, e.g., SDSS *ugriz* optical colour-space [63]. Therefore, the "UV Excess" method, generally used for $z < 2$

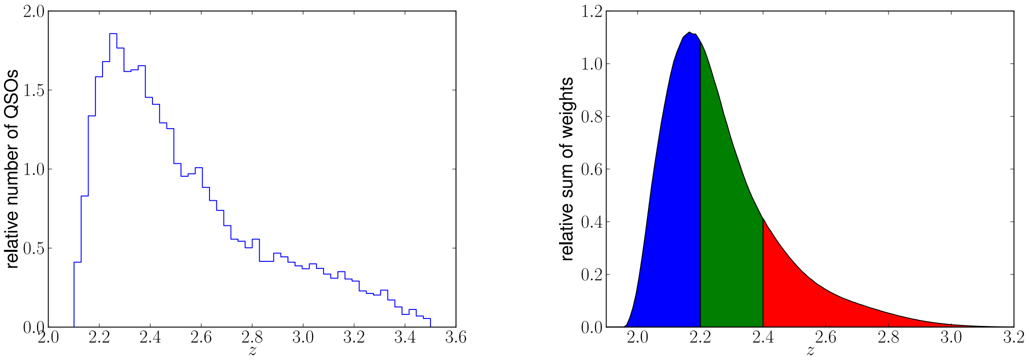

Figure 3. Redshift distribution of quasars for the sample used in this analysis (left panel), and weighted distribution of Lyman-α forest pixel redshifts for the three redshift bins considered in this paper (right panel). The quasar redshifts are cut to be between $2.1 < z < 3.5$. The pixel weights are limited by the UV coverage of the spectroscope at low redshift end and by the redshift of quasars at high redshift. We show the sample of quasars without DLA flag; however, the two plots are virtually identical for the sample in which DLA flagged quasars are included.

objects, fails. As such, the $2.2 < z < 3.5$ redshift range was sparsely sampled in the original SDSS, and the number density of $z > 3.5$ objects was very low ($\sim$ few deg^{-2}).

In BOSS, selecting $2.2 < z < 3.5$ is key to meeting our science requirements. We have the advantage that when one selects the quasars only as backlights for the detection of neutral hydrogen, one is at liberty to use a combination of methods to select quasars without having to understand the completeness of the sample in any variable relevant for studying quasar properties or clustering. In total 54,909 spectra were taken by BOSS that were targeted as high, $z > 2.2$, quasar targets between Modified Julian Dates (MJDs) 55176 (11 Dec 2009) and 55383 (6 July 2010). From this sample, 52,238 are unique objects, and have SDSS PSF magnitudes in the range $18.0 < r < 22.0$. Of these 52,238 objects, 13,580 (14,810) quasars have $z \geq 2.20$ (2.10) and a secure redshift as given by the BOSS spectro-pipeline.

Although the BOSS spectro-pipeline is secure at the 90-95% level, a non-negligible number of objects are classified as QSOs when they are stars and vice-versa. Because of this, the nearly 55,000 objects were also visually inspected by one, or more, of the BOSS team members. These inspections are generally referred to as the FPG inspections. The manual inspections of the BOSS spectra give a further level of confidence in the data, and result in a net gain of real high-z quasars that the pipeline does not select, at the level of $\sim 1-3$ objects per 7 deg^2 plate. Moreover, the visual inspection has the option to manually label high-z quasars which have either Damped Lyman-α (DLA) or Broad Absorption Line (BAL) features. Currently, this option is binary in the sense that the visual inspector has the option of setting the BAL and DLA flags but no attempt is made to measure $N_{\rm HI}$ column density or the position of the absorber. We have checked that the flagged DLAs have $\log N(\text{H I})(\text{cm}^{-2}) >\sim 20.3$. The list is not complete however, since DLAs are difficult to detect along lines of sight of SNR < 4 [64]. We have also checked that the so-called balnicity index is larger than 2000 km s^{-1} (see e.g. [65]). The removal of DLA systems in our data hence still depends on a human factor and signal-to-noise ratio and is currently not a very well quantified process. This will become better defined in future work.

For the analysis presented here, those objects labeled by the FPG visual inspections as “BAL” are not used, and we don’t include objects labeled as “DLA” unless mentioned specifically.

In our sample we restrict the analysis to quasars from the FPG sample with $2.1 < z < 3.5$. Compared to the targeted science sample described above, this sample has no magnitude cut, expands the lower redshift to $z = 2.1$ and uses some quasars that are on plates nominally marked as “bad” by the pipeline, but contain useful data based on manual inspections. After removing approximately 1300 quasars marked as BALs, our complete sample contains 14,598 quasars, of which 13,743 are not labeled as having a DLA. Several quasars in the first year sample had repeated observations - we do not use these repeats and limit ourselves to the best quality observation (as determined by FPG). The FPG visual inspections are continuously being updated and revised. We froze the data used on 1st October 2010.

Further details about the target selection and details about the instrument and the observational procedure can be found in [59].

Figure 1 shows the position of the BOSS Year One quasars on the sky in equatorial projections in the Aitoff projection. Figure 2 shows the geometry of BOSS probes of large scale structure in comoving coordinates. Figure 3 gives the redshift distribution of the $z > 2.1$ quasars and the weighted histogram of pixel redshifts contributing to our correlation measurement (explained and discussed later), as a function of redshift.

3 Synthetic data

To test our data reduction pipeline and to examine the impact of various systematics, we created extensive synthetic datasets. These synthetic data shared the geometry with the observed set of quasars, i.e. the positions and redshifts of quasars were identical, and we used the observed SDSS g magnitudes to normalize the spectra.

The statistical properties of synthetic data matched our initial estimate of the properties of the underlying flux field, including the non-linear corrections to the linear redshift-space distortions [39].

The Lyman-α absorption spectra of each quasar were generated using a method described below to obtain the variable $\delta_F = F/\bar{F} - 1$, where F is the fraction of transmitted flux, at each spectral pixel from a realization of the underlying cosmological density fluctuations, including linear redshift-space distortions and non-linear corrections to the flux power spectrum model that is used as input. A total of 30 different realizations of this underlying cosmological field were generated. For each realization, the following sets of synthetic data were created, with increasing level of realism to evaluate the impact of various observational effects and of our data reduction procedure:

- Noiseless measurements of δ_F
- Noiseless measurements of Lyman-α forest with continua
- Noisy measurements of Lyman-α forest with continua
- Noisy measurements of Lyman-α forest with continua and high-absorption systems
- Noisy measurements of Lyman-α forest with continua and forest metal contamination
- Noisy measurements of Lyman-α forest with continua and forest metal contamination and high-absorption systems

Note that continua here means variable continua using randomly generated PCA eigenmode amplitudes [66], instead of the same mean continuum for every QSO.

In the following subsections we briefly describe how these synthetic data are actually generated. The generation of Lyman-α absorption and associated structure will be discussed in more detail in [67].

3.1 Absorption field

3.1.1 Generation of a Gaussian field

A Lyman-α forest survey samples the intrinsically three-dimensional flux transmission fraction field along a set of infinitesimally thin lines of sight to quasars. The brute force method for generating a survey with correct correlations would be to generate the full 3D field with sufficient resolution (tens of kpc), and then read off the density along lines of sight. For a survey as large as ours, this requires too much memory. Fortunately, when generating a Gaussian field we can greatly simplify the problem, maintaining exact high-resolution statistics while only creating the field for the limited fraction of the volume where it is required.

To generate a Gaussian random field for a general set of pixels with covariance matrix $\mathbf{C}$ we can Cholesky decompose the covariance matrix, i.e. first find $\mathbf{L}$ such that

$$\mathbf{L} \cdot \mathbf{L}^T = \mathbf{C}. \tag{3.1}$$

If we then generate a unit variance white noise field w_j in the pixels and multiply it by L_{ij}, the resulting field will have the desired correlation function, i.e.

$$\langle \delta_i \delta_j \rangle = \langle L_{ik} w_k L_{jl} w_l \rangle = L_{ik} L^T_{kj} = C_{ij}. \tag{3.2}$$

Assuming the lines of sight are parallel allows another significant simplification. Suppose the field $\delta\left(x_\parallel, \mathbf{x}_\perp\right)$ has power spectrum $P\left(k_\parallel, k_\perp\right)$. If we Fourier transform this field in the radial direction only, the resulting modes have correlation

$$\left\langle \delta\left(k_\parallel, \mathbf{x}_\perp\right) \delta\left(k'_\parallel, \mathbf{x}_\perp{}'\right) \right\rangle = 2\pi\ \delta^D\left(k_\parallel + k'_\parallel\right) P_\times\left(k_\parallel, \left|\mathbf{x}_\perp - \mathbf{x}_\perp{}'\right|\right)\ , \tag{3.3}$$

where

$$P_\times\left(k_\parallel, r_\perp\right) = \frac{1}{2\pi} \int_{k_\parallel}^\infty k\ dk\ J_0\left(k_\perp r_\perp\right)\ P\left(k, \mu_k\right)\ , \tag{3.4}$$

where $k_\perp = \sqrt{k^2 - k_\parallel^2}$ and $\mu_k = k_\parallel / k$. The key point is that modes with different values of $k_\parallel$ are uncorrelated. We can generate the field efficiently by following the above general procedure for generating a correlated Gaussian random field, except now independently for every value of $k_\parallel$ required for the radial Fourier transform. We never need to manipulate a matrix larger than a manageable $N_q \times N_q$, where N_q is the number of quasars. However, we do take into account the fact that lines of sight are not fully parallel as discussed in Section 3.1.3.

We use this procedure to generate any desired Gaussian field δ at each pixel of each quasar spectrum of the synthetic data sets, once we have a model for the power spectrum $P(k, \mu_k)$ of this variable.

3.1.2 Generation of the transmitted flux fraction, F

In order to obtain realistic synthetic data including noise as in the observed spectra, it is necessary to generate the flux transmission fraction F with a realistic distribution in the range 0 to 1, and add the noise to this flux variable. The solution we have adopted is to convert the Gaussian fields δ to a new field $F(\delta)$ with any desired distribution. In this paper we use a log-normal distribution in the optical depth, $\tau = -\log F$ (where log denotes natural logarithm), which implies the following probability distribution for F:

$$p(F)\,dF = \frac{\exp\left[-(\log\tau - \log\tau_0)^2/(2\sigma_\tau^2)\right]}{\tau F\sqrt{2\pi}\sigma_\tau}\,dF \ , \tag{3.5}$$

with the two parameters τ_0 and σ_τ that determine the mean and dispersion of F. We find the function $F(\delta)$ that results in this distribution for F when δ is a Gaussian variable. The correlation function of the variables δ and F are then related by the following equation:

$$\begin{aligned}
\xi_F(r_{12}) &= \langle F_1 F_2\rangle \\
&= \int_0^1 dF_1 \int_0^1 dF_2 p(F_1,F_2)F_1F_2 \\
&= \int_{-\infty}^{\infty} d\delta_1 \int_{-\infty}^{\infty} d\delta_2 p(\delta_1,\delta_2)F_1F_2 \\
&= \int_{-\infty}^{\infty} d\delta_1 \int_{-\infty}^{\infty} d\delta_2 \frac{e^{-\frac{\delta_1^2+\delta_2^2-2\delta_1\delta_2\xi^2(r_{12})}{2(1-\xi^2(r_{12}))}}}{2\pi\sqrt{1-\xi^2(r_{12})}} F(\delta_1)F(\delta_2) \ .
\end{aligned} \tag{3.6}$$

We note that this equation relates ξ_F to ξ, independently of the variables r, μ and z. Therefore, we simply tabulate $\xi_F(\xi)$, and invert the relation to figure out the correlation function ξ that is required in order to obtain any desired correlation function ξ_F.

We use a model for the flux power spectrum in redshift space, $P_F(k,\mu_k)$, which was fitted to the results of simulations in [39]. We use the value of β for the central model in that paper, $\beta = 1.58$, and a bias inspired by the central model but roughly adjusted for cosmological model (not using the parameter dependence Table in [39]), $b = 0.145$, at $z = 2.25$. When generating the synthetic data, we keep the value of β constant as a function of redshift, and vary the amplitude of the power spectrum according to a power-law, $P_F \propto (1+z)^\alpha$. We compute the correlation function ξ_F from the Fourier transform, then we calculate ξ for the Gaussian field, and we Fourier transform back to obtain the desired power spectrum for δ. We generate this Gaussian field in all the quasar spectra with the method described above, and then transform this to the variable δ_F. In this way we obtain a set of spectra with the desired flux distribution and redshift-space power spectrum.

We use mean transmission fraction approximately matching the observations [8]: $\ln\langle F\rangle(z) = \ln(0.8)\left[(1+z)/3.25\right]^{3.2}$.

3.1.3 Redshift evolution and non-parallel lines of sight

The field δ_F is generated as described at a fixed redshift, and assuming parallel lines of sight. This is done at four different values of the redshift, by generating the realizations of the field with fixed random numbers for all the Fourier modes, and changing the amplitudes according to the different power spectra at every redshift. The power spectrum varies both because of the intrinsic evolution, and because the angular diameter distance used to convert angular to physical separation between the quasars varies with redshift. In this way, the fact that the lines of sight are non-parallel is incorporated into the calculation in the same way as the redshift evolution of the power spectrum. The final field δ_F is obtained by interpolating between the fields that have been generated at different redshifts, to the exact redshift of each point on the line of sight. This field then has the desired local power spectrum at any redshift, evolving as it is prescribed according to the linear interpolation.

The redshift evolution of amplitude assumed in this work is described by a power law $(1+z)^\alpha$, with $\alpha = 3.9$ as suggested by the measured evolution of the 1D flux power spectrum [9].

3.2 Adding absorption systems: Lyman Limit Systems (LLS) and Damped Lyman-α (DLA) systems

The highest column density absorption systems produce broad damped absorption wings which can have a strong impact on the correlation function of the transmitted flux. These systems are traditionally known as Damped Lyman Alpha absorbers (DLA) and have H I column densities above $10^{20.3}\mathrm{cm}^{-2}$, however, the damping wings can affect the line profile at lower column densities as well. Systems with column densities above $10^{17.2}\mathrm{cm}^{-2}$ are known as Lyman Limit Systems (LLS) since they are self-shielded [68]. At $10^{17.2}\mathrm{cm}^{-2}$, the effect of damping wings on the profile is small, but it becomes significant well before $10^{20.3}\mathrm{cm}^{-2}$ [69].

The impact of these absorption systems is two-fold. First, they add noise to the measurement of the correlation function. Second, the systems trace the underlying mass density field, and therefore they are correlated with themselves and with the Lyman-α absorption arising from the intergalactic medium. This systematically modifies the overall Lyman-α transmission correlation function.

To simulate the effect of these systems in the synthetic data, we introduce lines with $N_{\mathrm{HI}} > 10^{17.2}\mathrm{cm}^2$ with an abundance consistent with observational constraints [70–72], using the formula of [69]. The decrease in $\langle F\rangle\,(z = 2.25)$ due to these systems is 0.014. We also introduce a correlation between these systems and the rest of the Lyman-α absorption by placing them only in regions where the optical depth is above a critical value τ_0, such that the probability of $\tau > \tau_0$ is 1%. This is performed to explore the way that the damped absorbers may bias the measured correlation function, but their detailed impact depends on the specifics of this correlation. We leave for a future analysis a better modeling of the cross-correlation of damped absorbers and the Lyman-α forest.

In the rest of the paper we refer to these contaminants as LLS/DLA.

3.3 Adding forest metal contamination

Absorption by intergalactic metals imprints correlated signal in quasar spectra on characteristic scales. These scales are set by the wavelength ratios of metal line transitions with Lyman-α and with each other. As a result, this correlated signal is a potential contaminant of large-scale structure measurements in the forest. In order to add forest metal absorption to the synthetic data we measure the strength of metals. We do this in a self-consistent manner by measuring metal line absorption in the continuum normalized BOSS Year One spectra (excluding spectra with known DLAs). We use a modified version of the method set out by [73] and measure these lines by stacking absorption lines between 1041Å–1185Å in the quasar rest-frame. We measure the signal associated with metal lines correlated with

Lyman-α or Si III in the forest. This is described in Appendix C, including a look-up table of flux decrements measured in the composite spectra.

We introduce these metal absorption features to the full suite of mock data with no noise and no LLS/DLAs, on a pixel-by-pixel basis; we walk through the spectra and lower the transmitted flux by values from interpolation of this table, and then add Gaussian noise. As a result, we lower the mean transmitted flux by 0.003 in the mocks. This approach assumes that metal lines trace Lyman-α structure monotonically, and as such these 1D correlations will add metal structure to our 3D analysis. Full line profiles are recovered by virtue of our easing of the local minimum requirement (metal absorption is added at the wings of Lyman-α profiles as well as the center).

This technique does not provide a full description of metal absorption and, in particular, neglects the impact of scatter in metallicity and metal complexes. We test these mocks by stacking them in the same manner used for the BOSS Year One spectra. The metal correlations imprinted in the noise-free mocks are reasonably well recovered in these composite spectra. Composite spectra of the mocks with noise added (after metals have been introduced) show metal correlations that, where measurable, are up to 10% weaker than those seen in the observed data or added to the mocks in all cases except the Si III line seen in the strongest bin which is 30% weaker. This is caused by a combination of Gaussian noise and the probability distribution function of the flux. The noise distribution is symmetrical but, in the relevant regime, there are always far more pixels in the higher flux bin, which have weaker associated metal absorption. We conclude that metals added are probably an underestimate of the average metal absorption associated with Lyman-α lines, but these results are sufficient for an exploration of the approximate impact of forest metals. It seems unlikely that LLS/DLA interlopers are able to produce the observed metal signal for the reasons given in [73], however, we shall explore this issue further in future publications. We also combine forest metals with LLS/DLA corrected mocks by introducing these high column lines after forest metals have been added. Hence only metals associated with the Lyman-α forest are included.

3.4 Generating the spectra

Once we have created an absorption field for every line of sight, we proceed to generate the actual spectrum for each quasar, multiplying it by the "continuum" of the quasar, i.e. the unabsorbed spectrum.

We generate each quasar continuum shape taking into account the quasar redshift and using a mean rest-frame continuum and random PCA components derived from the low-redshift Hubble data [66]. The continuum is then normalized using the g magnitude of the quasar (taking into account the Lyman-α forest absorption).

Since our data are sampled on precisely the same grid as the observed data, we can also introduce noise by using the actual values of noise from the observed data. We assume noise to be Gaussian with the absolute flux variance given by the pipeline, i.e., we do not correct the signal-to-noise ratio for differences between our randomly generated continuum level and the data level.

Because the mocks were generated before the data analysis procedure was finalized, we have mocks only for quasars with redshift > 2.2, while the data analysis uses $z_q > 2.1$.

4 Data analysis

In this section we describe the analysis applied to both real and synthetic data. Briefly, the steps involved in this analysis start with co-adding the multiple spectra of each quasar. We then fit a model for the mean quasar continuum and mean absorption from the whole set of spectra. This is used to determine the fluctuation in the fraction of transmitted flux, δ_F, from the observed flux, in each individual spectrum over the Lyman-α forest range. The correlation function is then measured from these flux fluctuations. The information on the distribution of datapoint pairs and the measured correlations feed into the code that estimates the errors on our correlation function. With the estimated correlation function and error-covariance matrix, we finally proceed to estimate the parameters in our model of the correlation function, in particular the two bias parameters. The next subsections explain in detail each of these steps.

4.1 Preparing the data

The data analysis from raw CCD images to reduced individual exposures is performed using the standard SDSS `spectro2d` and `spectro1d` pipelines version `v5_4_14`. Typically there are 3 to 5 exposures of each quasar that need to be co-added. We do not use the co-added data from the pipeline. Instead, we co-add the data using the same fixed wavelength grid as the pipeline, but combining the flux values of the closest pixels in wavelength from the individual exposures. This ensures, in the simplest possible way, that the noise of the co-added data is independent from pixel to pixel, at the expense of a poor treatment of the small scale fluctuations in the data. Since we are interested in large-scale correlations, we are not concerned about these effects. In each pixel the data are coadded using inverse variance weighting. We apply a small correction to the final variance in each pixel, which typically increases it by less than 10%, to ensure that the inter-exposure variance in pixels is consistent with the noise predicted by the pipeline for individual exposures.

An example of a BOSS spectrum after this reduction is shown in Figure 4. This spectrum has a somewhat higher than typical SNR of quasars in our sample.

4.2 Continuum fitting

The next step in analyzing the data is to fit the continuum. Our model for the flux measured at wavelength λ in quasar i at redshift z_i is as follows:

$$f(\lambda, i) = a_i \left[\lambda_r/(1185\,\text{Å})\right]^{b_i} C(\lambda_r)\bar{F}(\lambda)\left[1 + \delta_F(\lambda, i)\right] \ , \tag{4.1}$$

where $\lambda_r = \lambda/(1 + z_i)$ is the rest-frame wavelength. The term $C(\lambda_r)$ denotes the mean rest-frame quasar spectrum, which is multiplied by a power law $a_i[\lambda/(1185\,\text{Å})]^{b_i}$, where a_i and b_i are two parameters determined for each individual quasar. The power-law is included to allow for large-scale spectro-photometric errors that are present due to imperfect sky subtraction and calibration, as well as for any intrinsic variation in the quasar spectra. The term $\bar{F}(z)$ describes the mean absorption in the forest as a function of redshift. The entire product $a_i\lambda^{b_i}C(\lambda/(1 + z_q))\bar{F}(z)$ is a well constrained quantity, but individual terms are correlated. For example, we can multiply a_i and $\bar{F}(z)$ by a certain factor and absorb it into $C(\lambda/(1+z_q))$; this is also true for the power-law parameter b_i. Consequently, some quantities ($\bar{F}$, a_i, etc.) are determined only up to an overall normalization constant.

For the purpose of determining our fit to the mean quasar continuum and mean transmission, and the parameters a_i and b_i, we calculate an overall likelihood of all the data

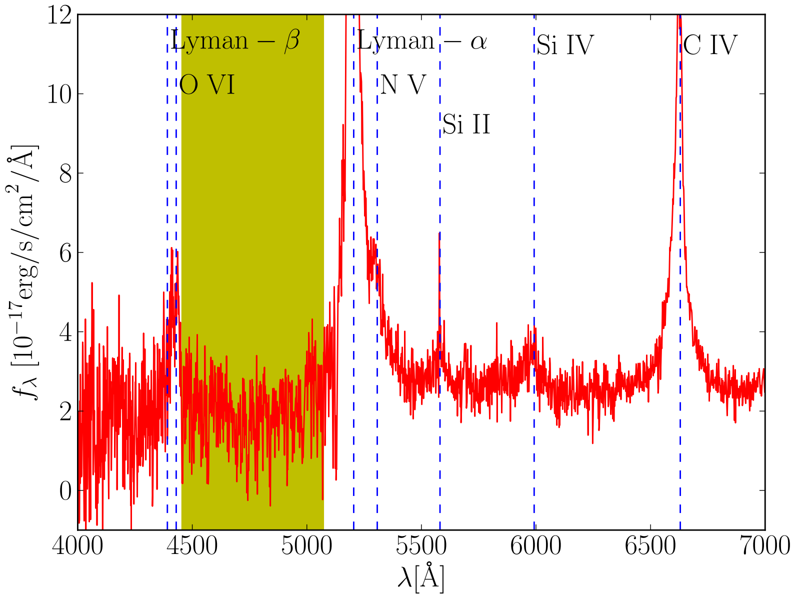

Figure 4. A quasar spectrum at redshift $z = 3.276$ from the first year of BOSS data. The Lyman-α forest is the shaded region between the Lyman-α and the Lyman-β emission lines. Other strong emission lines are also indicated.

without taking into account the correlations of the forest and the continuum fluctuations in different spectral pixels. This simplification allows us to write this simple expression for the likelihood function,

$$\log \mathcal{L} = \sum_i \sum_\lambda \left[-\frac{\left[f(\lambda, i) - a_i[\lambda_r/(1185\,\text{Å})]^{b_i}\, C(\lambda_r)\bar{F}(z)\right]^2}{2\sigma^2(i,\lambda)} - \log\sigma(i,\lambda) \right] + \text{const.}. \quad (4.2)$$

Here, $\sigma(i, \lambda)$ is the sum of the variances due to measurement noise and the *intrinsic variance*, i.e., it contains both variance due to continuum errors as well a variance due to small scale flux fluctuations. This quantity is required to be a function of the observed wavelength only in the region of the Lyman-α forest (or equivalently, the Lyman-α forest redshift), while outside the forest we force it to depend only on rest-frame wavelength, which implicitly assumes that the dominant source of variation in that region is continuum errors. We do not necessarily believe his assumption is true, but this allows in one sense optimal weighting of different parts of the quasar spectrum outside the forest, i.e., prevents certain high-variance parts of the rest-frame spectrum from excessively affecting the solution. In the long term, this assumption will have to be relaxed.

We restrict the rest-frame wavelength range used for this fit between 1041 Å and 1600 Å, and we also discard the data at an observed wavelength bluer than 3600 Å. We parameterize $C(\lambda_r)$ using 20 equally-spaced spline nodes inside the forest (1041 Å to 1185 Å), and 20 equally-spaced spline nodes outside the forest (1185 Å to 1600 Å), with one point fixed to the center of the Lyman-α emission line. We also use 8 spline nodes in redshift for the mean transmission fraction $\bar{F}$, another 8 spline nodes to describe the scatter in the forest as a function of redshift, and another 20 spline nodes to describe the rms variations with rest-frame wavelength outside the forest. The full model used in continuum fitting is therefore described by 76 global parameters and 2 parameters for each quasar, a_i and b_i.

Since we are completely ignoring the correlations between pixels, the numerical values of $\sigma^2(i, \lambda)$ and $\chi^2 = -2 \log \mathcal{L}$ have no physical interpretation. They should be thought of as merely assisting the fit to naturally incorporate the variances in the data. Therefore, it is not crucial that σ^2 is parameterized by redshift inside the forest and rest-wavelength outside. A possible criticism of our fitting procedure is that redward of the Lyman-alpha emission, 20 nodes of the spline fit are not enough to resolve the full structure of the mean continuum at the resolution of the BOSS spectrograph. However, these points are used only to assist with determining the values of a_i and b_i (the normalization and power law slope, relative to a mean continuum), and hence our determination of continuum must be only good enough to broadly describe the shape in that region.

We determine all the parameters using a specially crafted Markov Chain Monte Carlo (MCMC, see e.g. [74]) procedure to find the maximum of $\mathcal{L}$. The global parameters are first varied, while keeping a_i and b_i fixed, taking several hundred accepted samples (steps in the chain) of the MCMC. Next, the global parameters are fixed and we vary just a_i and b_i for one quasar at a time, looping over all quasars. At this step, only the model for the particular quasar under consideration matters, and hence likelihood evaluations are extremely fast. We then go back to varying the global parameters and repeat the process for several accepted samples. This is iterated about 50 times (we have tested that using just 20 iterations negligibly affects the final results), and we finally take the sample with the optimal fit. The proposal function of the MCMC process learns the degeneracy directions by using the covariance of previously accepted samples, separately for the global parameters and the quasar-specific parameters. This MCMC process is simply used here as a functional maximizer, i.e., we take just the most likely sample. The process takes about 36 hours on an 8-core workstation.

4.3 Determination of δ_F

The next step is determining the flux variation δ_F, reducing the resolution of our data and determining the weighting factors that are used for calculating the correlation function. First, the continuum is predicted at each pixel of each quasar spectrum. Then the observed flux and continuum are averaged over four consecutive pixels, weighted with the inverse variance. This provides coarser pixels and an estimate of the noise on the value of flux in each rebinned pixel, with a reduced resolution. The typical length of these rebinned pixels is $\sim 3h^{-1}$Mpc. We assume that the total variance in a coarse pixel is the sum of the noise plus the intrinsic variance in the flux field, which is a function of redshift. We then iteratively determine:

- The intrinsic variance in the rebinned pixels, determined as a function of redshift by splining over 8 points in redshift.
- A factor by which we multiply our estimate of the continuum in each quasar, which ensures that the flux averaged over the entire Lyman-α forest of the quasar equals the mean continuum averaged over the same interval. This average is calculated by weighting with the inverse total variance in each pixel.

We find that by forcing the average mean flux over the quasar forest to match that of the average continuum, we decrease the variance in the correlation function estimate, although we also erase all modes along the line of sight of wavelength larger than that of the size of the forest. The way that this elimination of the mean flux variation in each quasar alters the measured correlation function can be easily modeled, as discussed in Appendix A. The upshot is that we can correct the predicted correlation function for this effect using a simple theory with one parameter Δr, which is the effective Lyman-α forest length. In the limit of $\Delta r \to \infty$, the correction disappears, as expected. The value of Δr can be predicted from first principles, but we let it be a nuisance parameter as explained later.

Finally, we find the average value of the fluctuation δ_F at each fixed observed wavelength by averaging over all quasars, and we then subtract this average from all the values of δ_F in each individual quasar. The average is calculated in wavelength intervals $\Delta\lambda/\lambda_{Ly\alpha} \ = 0.001$. This procedure, which removes purely radial modes, helps remove some systematics of imperfect sky subtraction or Galactic CaII absorption [75] that occur at a fixed wavelength. The analysis of synthetic data demonstrates that this negligibly impacts the measured Lyman-α correlation properties. Finally, we make a cut on pixels in δ_F that are over $6-\sigma$ away from zero in terms of the total variance. A handful of pixels were removed that way, but this cut makes negligible change in the resulting correlation function.

4.4 Estimation of the correlation function and its errors

We use the trivial sub-optimal estimator of the correlation function as simply the weighted average over pixel pairs,

$$\bar{\xi}_F(r,\mu) = \frac{\sum_{\text{pairs } i,j} w_i w_j \, \delta_{Fi}\delta_{Fj}}{\sum_{\text{pairs } i,j} w_i w_j}, \tag{4.3}$$

where the weights w_i are the total inverse variance (from both the measurement noise as well as the intrinsic variance due to small scale fluctuations in the forest). This estimator effectively assumes that the covariance matrix of the δ_{Fi} values can be approximated as diagonal. By construction it is an unbiased estimator, but it is not an optimal one. To avoid introducing contamination from the correlated residuals from continuum fitting errors, we include only pairs of points from *different quasars* when measuring the three-dimensional correlation function. We have also found that including pairs of pixels from the same observed-frame wavelength produces strong contamination, presumably from the residuals of the sky-subtraction, and we therefore eliminate all pairs of pixels with an observed wavelength that are within 1.5 Å of each other. We also measure the one-dimensional correlation function along each quasar, which is then used for the error estimation.

We measure the correlation function $\xi_F(r,\mu,z)$ in 12 radial bins up to $r = 100h^{-1}$Mpc, 10 angular bins equally spaced in $\mu = \cos\theta$, where θ is the angle from the line of sight, and 3 redshift bins ($z < 2.2$, $2.2 < z < 2.4$, $z > 2.4$). The redshift bins correspond to the redshift of the absorbing gas rather that the redshift of the background quasars backlighting the gas. Together with estimating the correlation function in individual bins, we have also calculated the weighted averages of r, μ and z for each bin (i.e., weighted in the same way as the ξ measurement in that bin), which are used in the subsequent analyses. These averages correspond to the true bin position, although we find the difference with respect to the nominal bin centre to be small. When estimating the correlation function in limited redshift bins, we take into consideration all pairs whose mean redshift falls in the redshift bin under consideration.

It can be shown that the error covariance matrix of this estimator is given by (see Appendix B):

$$C_{AB} = \frac{\sum_{\text{pairs } i,j\in A,\text{pairs } k,l\in B} w_i w_j w_k w_l (\xi_{Fik}\xi_{Fjl} + \xi_{Fil}\xi_{Fjk})}{\sum_{\text{pairs } i,j\in A} w_i w_j \sum_{\text{pairs } k,l\in B} w_k w_l} \, , \tag{4.4}$$

where A and B represent two bins in r, μ and z of the correlation function measurement and $\xi_{ij,obs}$ denotes the actual observed covariance between the data points i and j. We stress that $\xi_{ij,obs}$ is obtained from the data, so the contribution from noise and continuum fitting errors, metal absorption or any other possible systematic effects is automatically included (note that we do use overall measurements of the correlation function, including even the noise in an averaged sense, not pixel-by-pixel noise, however, the covariance is not dominated by pixel self-products so this detail probably is not important).

Strictly speaking, this estimator for errors is true only at the level of 2-point contribution to the error covariance matrix; however, we show using synthetic data that it accurately reproduces the errors (although our mocks do not contain as much non-Gaussianity as we expect from the real data).

Evaluating the sum in the numerator of equation 4.4 is a computationally daunting task: one would need to make a sum over all possible pairs of pairs, which is an $O(N^4)$ task for N datapoints compared to the $O(N^2)$ for the estimation of the correlation function itself. Since in our case $N \sim 10^5$, the extra effort is $O(10^{10})$ computationally more expensive when compared to estimation of the correlation function. However, only a small fraction of all possible 4-point configurations add significantly to the sum, namely those configurations of points (i, j, k, l) for which (i, k) and (j, l) are close together (so that the corresponding values of ξ_F are large), and for which the distance between the pairs (i, j) and (k, l) is in the relevant range over which we want to estimate the correlation function. We therefore use the following Monte-Carlo procedure:

1. For each quasar, identify the list of quasars that are closer than the largest distance scale on which we are attempting to measure the correlation function ($100h^{-1}$Mpc). We denote such quasar pairs *neighbors*. We additionally identify a subset of neighbours which are at distances closer than $r_{\rm cn} = 30h^{-1}$Mpc. Such quasar pairs are denoted *close neighbors*.

2. Select a random two pixels in the dataset, corresponding to *neighbouring* quasars A and B[1]. These two points constitute a pair (i, j), which is held fixed while we loop over all possible pairs of points (k, l). Pairs (k, l) are chosen from all possible pixels in close neighbors of A and close neighbours of B. For each such quadruplet and for the two possible pairing of points, determine which covariance matrix element they belong to and add to the corresponding element according to the Equation (4.4).

3. Repeat the step 2 for $N_{\rm MC}$ times and then multiply the final sum with the ratio of all possible pairs (i, j) to the actual number of pairs considered $N_{\rm MC}$.

4. Divide this sum by the appropriate sum of weights for each covariance matrix bin.

[1] Since quasars have varying number of pixels, this is not identical to selecting random neighbouring quasars A and B and then random pixels in these quasars.

This process converges for about $N_{\rm MC} > 10^7$, in the sense that the resulting χ^2 values change by less than unity. The reason why this Monte-Carlo procedure works is fairly intuitive: quasars are distributed randomly and hence there are no special points that would anomalously contribute to the variance of our estimator. We must sample the geometry of the distribution of our points well enough to cover all typical configurations, and by that time the error estimation converges. It is also important to note that our error estimation is complete in the sense that it is based on the measured two-point statistics in the field. For example, the continuum errors add a large scale contaminating signal along each quasar. These errors manifest themselves as larger correlations in the auto-correlation function of the quasar and ultimately result in larger errors on our measurements of the three-dimensional correlation function.

Finally, the speed of this error estimation is drastically affected by the distance chose to denote close neighbors $r_{\rm cn}$. The method approximates that the contribution to the error covariance from correlations beyond this distance is negligible. The speed for converges grows approximately quadratically with $r_{\rm cn}$. We have tested the procedure with $r_{\rm cn} = 10, 20, 30, 50h^{-1}$Mpc and noted that the results converge at $r_{\rm cn} \sim 20h^{-1}$Mpc. The convergence has been established by inspecting the best fit χ^2 when fitting bias/beta parameters (see section 4.5) and demanding that it changes by less than unity. We have therefore chosen $r_{\rm cn} = 30h^{-1}$Mpc.

We tested this procedure using synthetic data as follows. We averaged 30 synthetic realizations of the full dataset (with noise and continuum) into one single mean measurement, which was assumed to be the true theory. For each realization we calculated the χ^2, resulting in 30 χ^2 values. Since we did not input the true theory, but the mean of thirty synthetic datasets, the mean χ^2, for N correlation function bins, is expected to be $\langle\chi^2\rangle = (1 - 1/M)\, N$, where $M = 30$ is the number of datasets being averaged over. In practice, we Monte-Carlo this distribution by drawing random vectors from the same covariance matrix in a set of 30. The purpose of performing error estimation test this way is that it allows us to disentangle the accuracy of error-estimation from systematic effects that might affect the theory predictions (for example, when continuum fitting can add small systematic shifts in the mean predicted signal). Figure 5 presents the results of this exercise on the synthetic data with continuum and noise and shows that the error estimation is robust.

We plot the structure of the error covariance in Figure 6 which shows that the neighboring μ bins are somewhat correlated, but the neighboring r bins much less so. The neighboring redshift bins (not plotted) are even less correlated (at around 1%). Finally, we note the negative correlation at high-μ, large r arising due to removal of the mean in each quasar.

The χ^2 distribution test in Figure 5 shows that the error calculation is broadly correct, but is not very sensitive to inaccuracy in an important subset of the error matrix, and does not test for systematic inaccuracy in the measurement. To address these points we also show the result of another test of the error matrix. In the next section we discuss how we fit the bias parameters, and in the results section we will quote the tail probability that β is larger than a certain number based on MCMC chains. We test this procedure on our 30 mocks by precisely the same test, namely fitting the bias parameters and then measuring the probability that the value of β is larger than the fiducial value used in creating the synthetic datasets. If our process is unbiased and correct, then we expect this probability to be uniformly distributed between zero and one. We plot the results of this test in Figure 7. Results of this exercise are quite encouraging - despite evidence for small systematic shifts in the inferred parameters (Table 1), we seem be able to successfully recover limits in the β

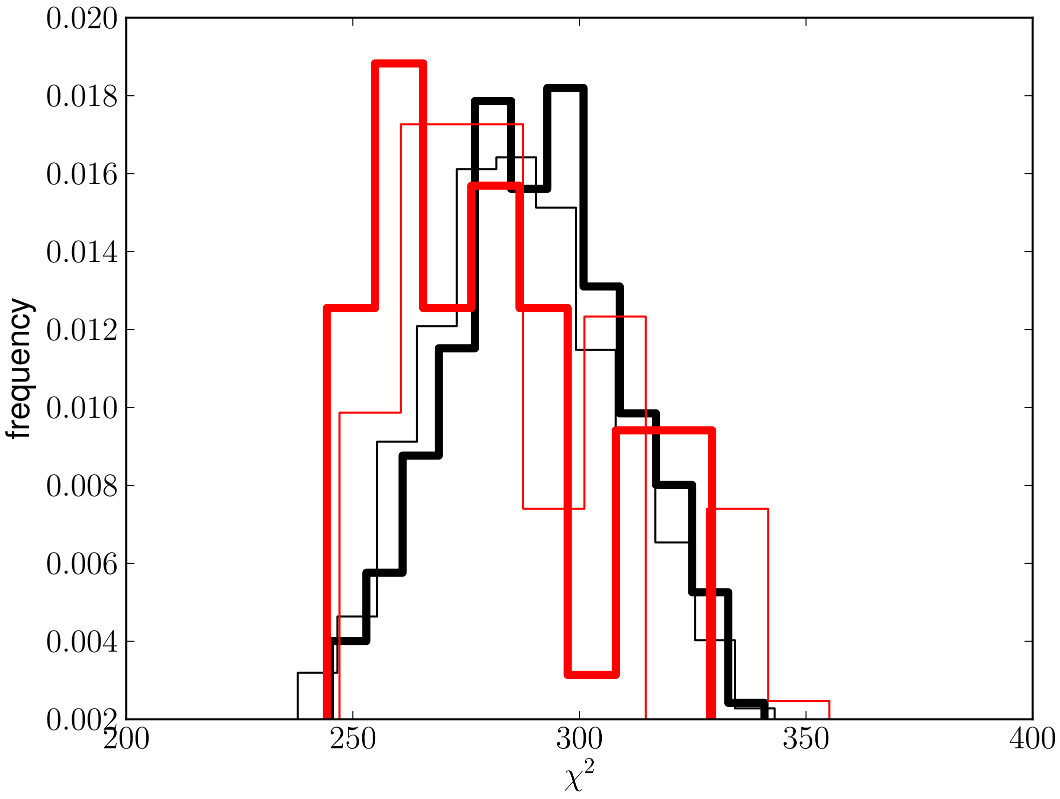

Figure 5. The result of the χ^2 distribution exercise for radial bins at $r > 10h^{-1}$Mpc, which is used in our cosmological fitting. The thick black histogram correspond to Monte-Carlos with realizations, while the thick red histogram is the distribution for the actual 30 realizations of synthetic data. The thin histograms of the same colors shows the effect of ignoring off-diagonal elements of the covariance matrix. The number of bins here is 300 ($10 \times 10 \times 3$). See text for discussion.

parameter for confidence limits used in this paper.

Note that Figure 7 is a test both of the errors (whether the covariance matrix is correct, and whether the likelihood is fully described by a covariance matrix) and any systematic offset in the measurement. There are a few things that are known to be imperfect about our analysis, although not expected to be important: We make no attempt to include non-Gaussianity of the measured field in the covariance matrix estimation, i.e., to include the connected part of the 4-point function. The mocks are in any case only a weak test of this assumption, as they are not designed to have the correct 4-point function. On the other hand, [9] found that the errors in the 1D power spectrum were within $\sim 10\%$ of the Gaussian expectation, and we are probing larger scales here, where we expect less non-Gaussianity. There are technical issues that could add up to biases in the pipeline. For example, the measured correlation function, which is used to estimate the errors, is assumed to be constant within radial bins, which is a poor approximation, especially for the most important bins at small separations. It is also true that the process of adding noise and fitting and dividing by continuum could potentially “renormalize” the large-scale bias parameters.

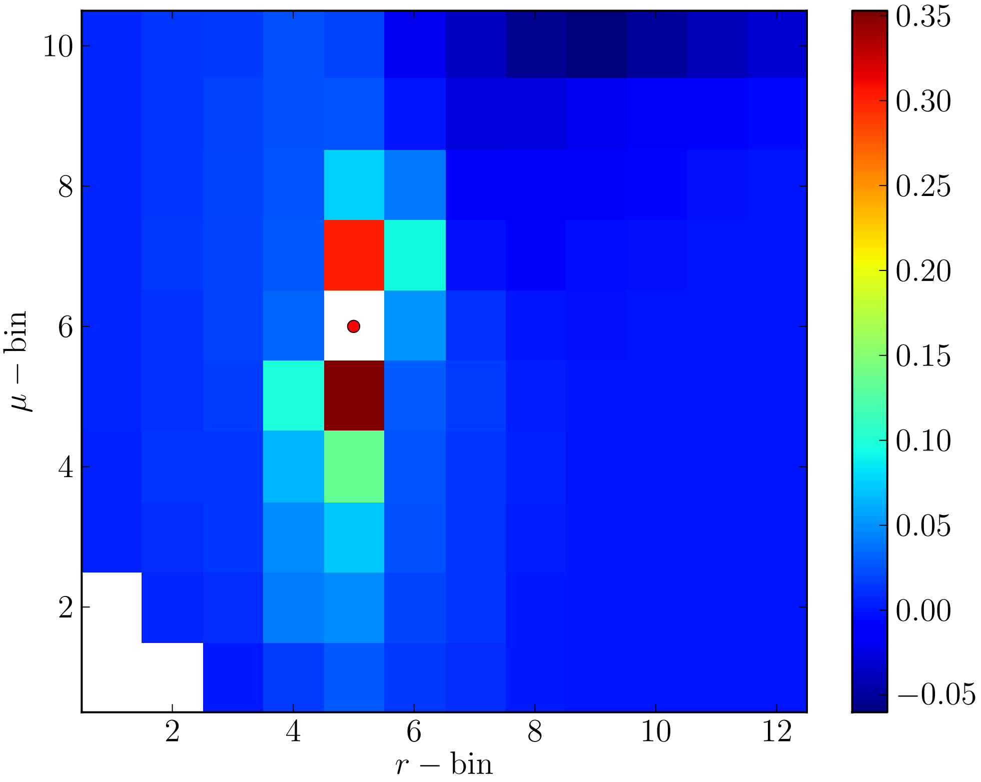

Figure 6. The structure of the error-covariance. Off-diagonal terms for the correlation matrix are plotted for one particular bin (denoted with a red point) on the plane of r and μ bins at the middle redshift bin. The covariance at the bin itself is unity, but we do not plot it to avoid unnecessarily stretching the color scale. The lower left corner is not present in the data due to a cut on observed-frame $\Delta\lambda$. Values of μ bins are linearly spaced between 0.05 (first bin) and 0.95 (tenth bin). The first four radial bins correspond to distances of $3h^{-1}$Mpc, $8h^{-1}$Mpc, $12h^{-1}$Mpc and $17h^{-1}$Mpc. The remaining 8 radial bins are uniformly spaced between $25h^{-1}$Mpc and $95h^{-1}$Mpc. The reference point is $25h^{-1}$Mpc, $\mu = 0.55$.

Careful exploration of these effects will be left for the future work. This will inevitably require many more mocks in order to clearly differentiate between biased error estimation and realization variance. In any case, Figure 7 shows that many possible forms of error in the analysis are unimportant.

4.5 Fitting bias parameters

We have fitted the bias parameters of the real and synthetic data assuming a fiducial flat cosmology with $\Omega_m = 0.27$, $\sigma_8 = 0.8$, $n_s = 0.96$, $H_0 = 71$ km s^{-1} Mpc^{-1} = 100 h km s^{-1} Mpc^{-1}, $\Omega_b = 0.04$, (where symbols have their standard meaning). We assume that the correlation function of δ_F is linearly biased with respect to the underlying dark-matter correlation function, and that it follows the linear theory redshift-space distortions. We do not use the measured correlation at $r < 10h^{-1}$Mpc for the fits where linear theory becomes a poorer ap-

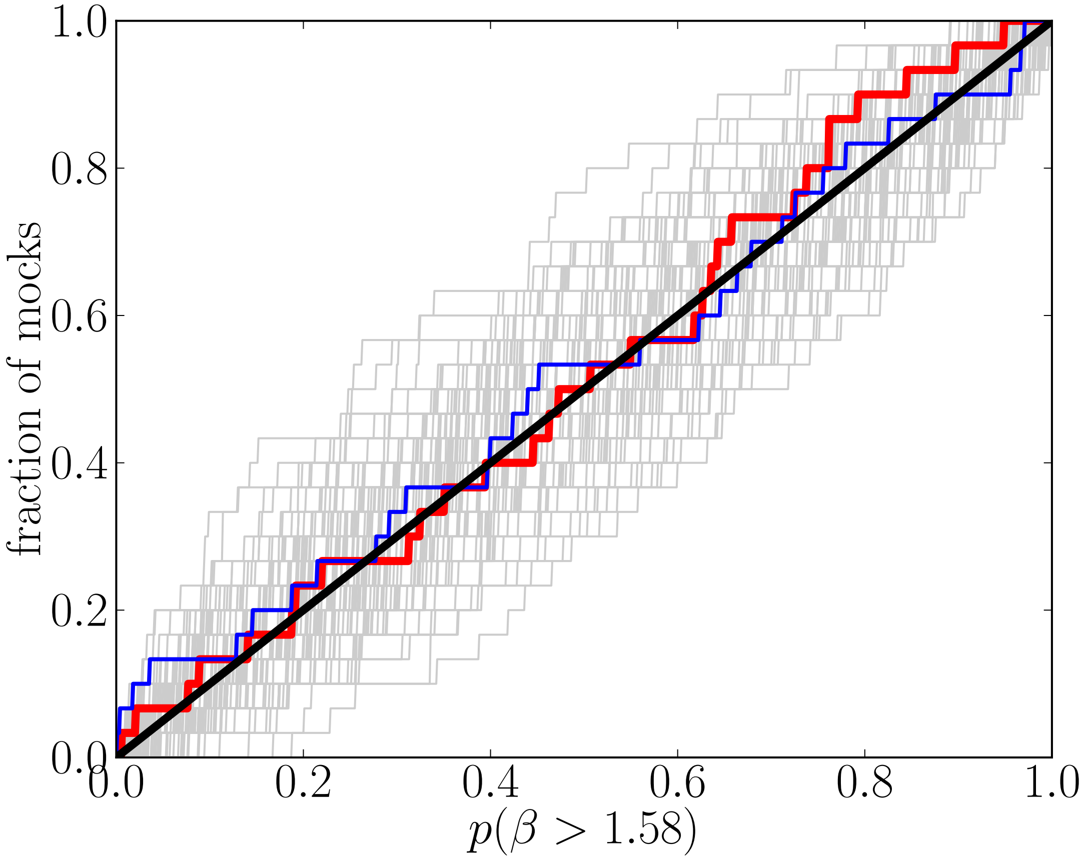

Figure 7. Cumulative distribution of the fraction of samples with value of β above the fiducial value when fitting synthetic data. For correctly estimated errors, ignoring the upper prior on β, this distribution should be flat between zero and unity, giving a straight line for cumulative distribution, plotted as thick red. Thin blue line is the same when off-diagonal elements in the covariance matrix are ignored. Faint grey lines are 100 realizations of 30 points drawn from uniform.

proximation. After Fourier transforming Equation (1.2), one derives the following expression for the redshift-space correlation function:

$$\xi_F(r,\mu) = \sum_{\ell=0,2,4} b^2 C_\ell(\beta)\xi_{F\ell}(r)P_\ell(\mu) \ , \tag{4.5}$$

where P_ℓ are Legendre polynomials,

$$C_0 = 1 + 2/3\beta + 1/5\beta^2 \ , \tag{4.6}$$

$$C_2 = 4/3\beta + 4/7\beta^2 \ , \tag{4.7}$$

$$C_4 = 8/35\beta^2 \ , \tag{4.8}$$

and

$$\xi_{F\ell}(r) = (2\pi)^{-3}\int P(k)k_\ell(kr)d^3k \ , \tag{4.9}$$

with the kernels $k_\ell(x)$ given by

$$\begin{aligned} k_0(x) &= \sin(x)/x \ , && (4.10)\\ k_2(x) &= (\sin(x)x^2 - 3\sin(x) + 3\cos(x)x)/x^3 \ , && (4.11)\\ k_4(x) &= (x^4\sin(x) - 45x^2\sin(x) + 105\sin(x) + 10x^3\cos(x) - 105x\cos(x))/x^5 \ , && (4.12) \end{aligned}$$

and $P(k)$ is the linear real-space power spectrum.

We model the redshift evolution of the power spectrum as a simple power-law,

$$P_F(k,z) = b^2 P(k, z=2.25)\left(\frac{1+z}{1+2.25}\right)^\alpha . \tag{4.13}$$

Therefore, the growth factor never enters our analysis - the redshift evolution is parametrised purely as power-law deviation from the power at $z = 2.25$. The only exception to this rule is when we fit the actual bias parameters in the three redshift bins independently (see Figure 20), where bias parameters are measured in individual redshift bins with respect to the matter power at that redshift (but are assumed constant across a redshift bin).

Whenever we fit for these parameters, we also include three Δr parameters describing the effect of removing the mean quasar component (as derived in Appendix A) at three redshifts, as free nuisance parameters. Although these parameters can be determined *a-priori*, we have found that fitting for them is easier, accurate enough (as tested on mocks), and worsens the errorbars on the inferred bias parameters by a very small amount when we marginalize over them. The resulting chains constrained values of Δr very weakly to be in the range between $\sim 200h^{-1}$Mpc and $500h^{-1}$Mpc.

We fit using an MCMC procedure. We put a flat prior on $0 < \beta < 5$ and unconstrained flat prior on $b(1+\beta)$, which are well-defined non-degenerate parameters. (Note that β can in general be less than zero, however, it is difficult to imagine a scenario in which this would happen.) This implies a non-flat prior on b. We also assume a flat prior on α.

When we fit the data, we always evaluate the theory at each bin's mean redshift, μ value and radius. This approach allows for any linear variations of the underlying quantity across the bin. This is not a good approximation at the lowest values of r, but, as we show later, it is a good approximation for the bins that we actually use when fitting the bias parameters.

4.6 Pipeline tests and synthetic data

Our code has been extensively tested on the synthetic data. These tests have served two goals. First, they helped remove coding errors in both the data-reduction code and the synthetic data making codes. For this purpose, the two codes were written completely separately by two people. Second, these tests established the amount of systematic effects introduced by the data-reduction process. Our guiding principle in this exploratory work was that we should only correct for the errors that affect our results at the level of the current statistical errors in the experiment. Namely, as long as we are able to reproduce the input parameters from our synthetic data with the highest level of realism and noise at acceptable χ^2 values, we pronounce the data reduction code sufficient and apply the same method to the observed data. More data might require better codes.

While not always strictly enforced, we did attempt to follow the principles of blind analysis: after our basic reduction code was in place and gave no visually surprising results on the observed data, we worked exclusively on synthetic data in order to understand the reduction procedure and returned to observed data only when we were able to reproduce the

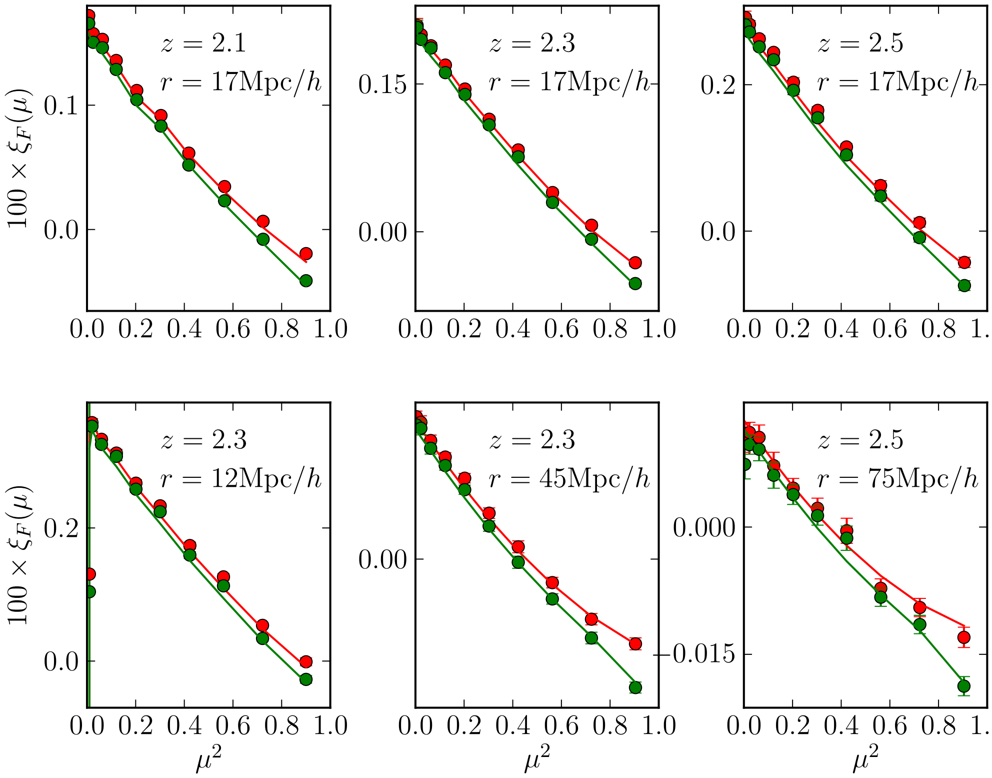

Figure 8. The correlation function plotted as a function of μ^2 for selected radial and redshift bins. Points are averages of 30 noiseless synthetic datasets with perfectly known continuum without any processing (red) and with mean and radial model removal (green). Red lines are the theory used to produce synthetic data, while green lines is the same after we corrected for the effect of mean removal using equations in Appendix A with Δr of 250, 300 and 350 h^{-1}Mpc for the three redshift bins.

input parameters from the synthetic data. (The analysis procedure was not, however, frozen at this point, as our mocks were not realistic enough to expect every problem to show up in them.)

We begin by discussing the noiseless synthetic measurements of δ_F. While these data are noiseless and there is no error associated with the continuum fitting, the final result still contains variance associated with sample variance.

We have analyzed these datasets using three methods: i) by calculating the correlation function, ii) by first removing the mean component of each spectrum and then calculating the correlation function and iii) by also removing the purely radial modes as described in Section 4.3. The result of ii) and iii) are practically identical. We show the results of this exercise in Figure 8, where we have averaged the individual results of all 30 realizations in order to make this an essentially statistical error-free measurement, at least when compared to the statistical error in the real data. The purpose of this figure is essentially two-fold: to illustrate that, at the level of noiseless synthetic data, we can reproduce the theory that is being put into the simulations, and, more importantly, that the simple model in Appendix

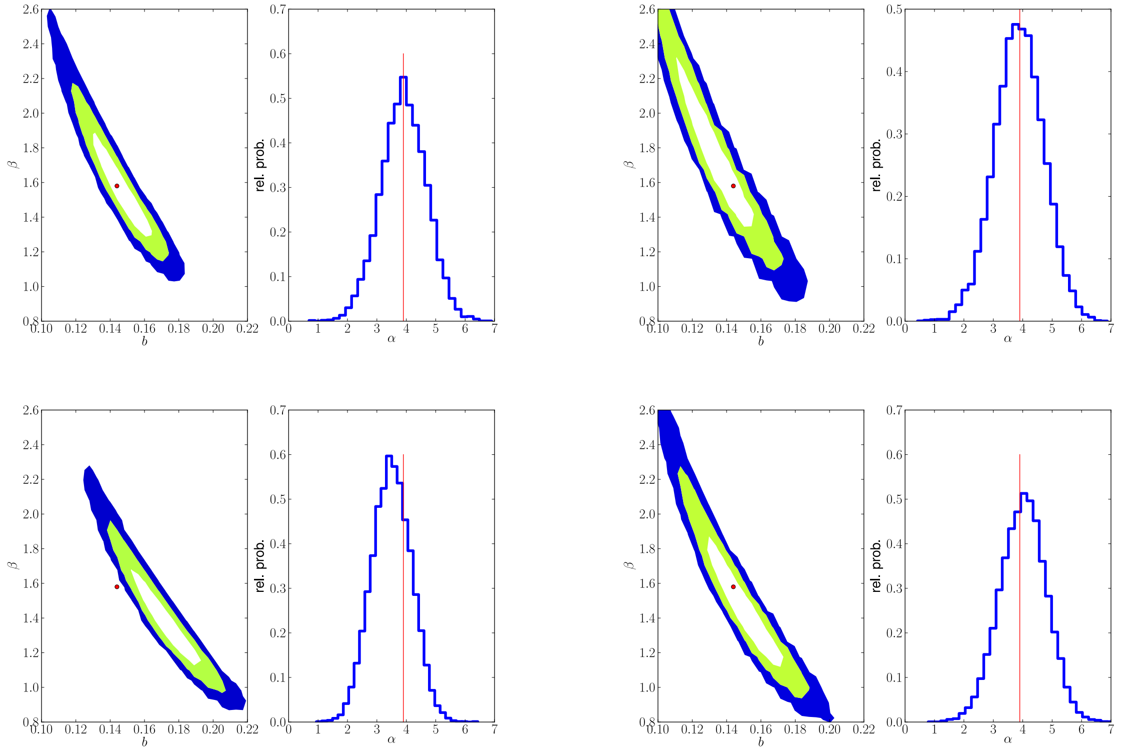

Figure 9. Results of fitting the data averaged over 30 mock datasets together with noise covariance for a single noisy realization and using only datapoints with $r > 20h^{-1}$Mpc in the fit. We show constraints on the $b-\beta$ plane and the probability histogram of α (which has negligible degeneracy with the other parameters). The input points are denoted by the red dot and the red line. The upper left plot is for the pure synthetic noiseless δ_F values. The upper right plot is for synthetic data that have PCA continua and noise. The lower left plot is for the data that in addition to PCA continua are additionally painted with high column-density systems. The bottom right panel is for synthetic data to which metals have been added as described in §3.3 (with noise and continua but no DLA/LSS).

A appears to produce fairly good results (we will quantify these later).

In order to proceed we need to understand the effect of adding various complexities to the data and how they affect the results. It is very difficult to judge the size of various effects by visually assessing how the points move on the plots, so we adopted the following procedure. We run the parameter estimation of the synthetic data with the covariance matrix corresponding to one mock noisy dataset, but with the data vector corresponding to mean over 30 mock datasets. During the fitting procedure this yields χ^2 values which are too small, but it allows one to see how central values move with respect to the size of the error bars. Note that our 3-parameter model (the bias b, the redshift-space distortion β, and the redshift-evolution index α) is essentially the same model that has been used to create mocks, but with the important difference that the latter contains corrections to the linear theory redshift-space formula arising from matter power spectrum non-linearities, scale-dependence of the effective bias and fingers-of-God effect.

In Figure 9 we show the results of the test described above, when fitting with the data corresponding to $r > 20h^{-1}$Mpc. Figure 10 contains the same material but for the sample

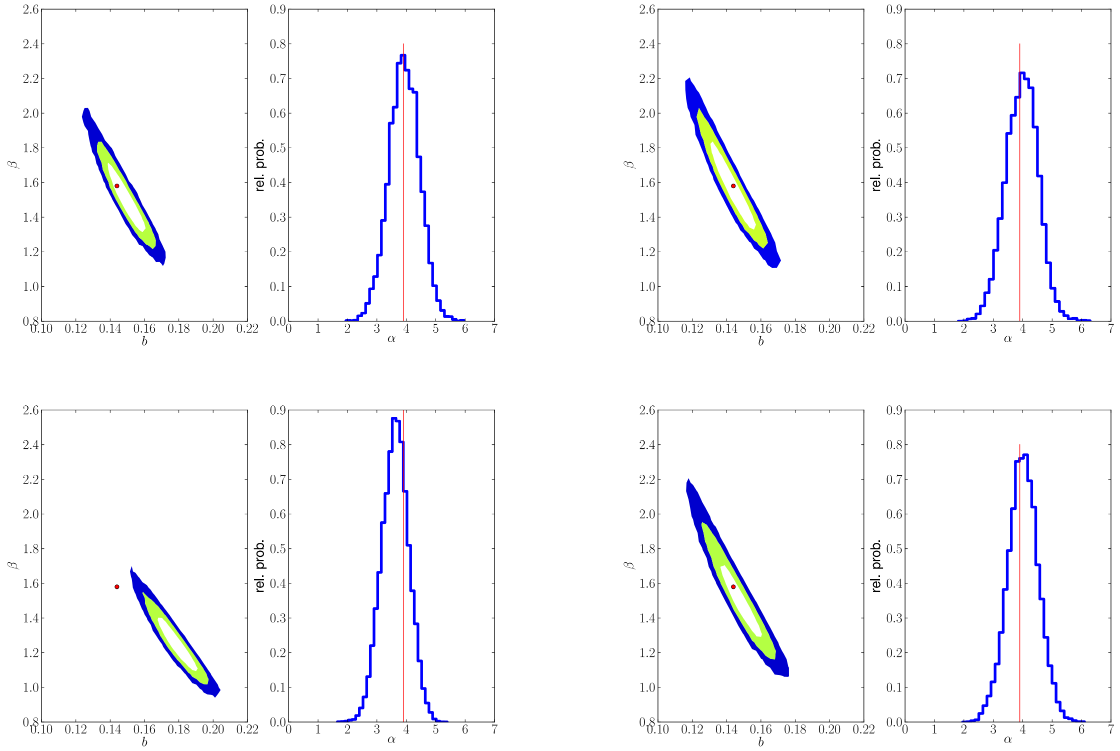

Figure 10. Same as figure 9 but for fitting all data with $r > 10h^{-1}$Mpc. The axis limits were kept the same for easier comparison.

for $r > 10h^{-1}$Mpc. We show the same information also in the tabular form in table 1. It is important to note that the inferred parameter constraints come with the usual caveats about Bayesian inference (projections and priors) and therefore we also quote the best-fit model, which should be at the position of the true theory, if we had infinite number of synthetic datasets.

The noise and continuum fitting do not strongly affect the inferred bias parameters. It is important to stress that this is a non-trivial conclusion: the continua in the synthetic data were simulated using PCA eigen-components, while the fitting procedure assumed a much simpler model for the continuum variations. This test also confirms the validity of the results of the Appendix A for the present purpose. There is some evidence that the continuum fitting may systematically decrease the bias, but the effect is smaller than the errorbars for one set of mocks. However, we find that the presence of the high column-density systems increases the expected bias and lowers the inferred value of β. This is not an artifact of the pipeline, but simply the result of change in the theory. Forest metals have the same effect, but to a somewhat lesser extent. High column-density systems and forest metals definitely warrant further investigation, but this is beyond the scope of this article, and here we just note that these two effects can affect the inferred values of bias, β, and, to a lesser extent, α. The inferred parameters are skewed (at a $0.5 - \sigma$ level) when fitting with linear models using $r > 10h^{-1}$Mpc data due to incorrect assumption of fully linear redshift-space distortions, especially at large values of μ.

Synthetic Data	bias	β	$b(1+\beta)$	α
Noiseless				
$r > 20h^{-1}$Mpc	0.145 ± 0.013	$1.58 \pm^{0.24\ 0.54\ 1.03}_{0.19\ 0.35\ 0.49}$	0.374 ± 0.008	3.93 ± 0.75
Best fit	0.148	1.53	0.375	3.95
$r > 10h^{-1}$Mpc	0.150 ± 0.007	$1.50 \pm^{0.13\ 0.28\ 0.43}_{0.11\ 0.22\ 0.32}$	0.373 ± 0.006	3.93 ± 0.54
Best fit	0.151	1.47	0.374	3.93
+ Continuum/noise				
$r > 20h^{-1}$Mpc	0.133 ± 0.017	$1.75 \pm^{0.39\ 0.93\ 1.83}_{0.29\ 0.52\ 0.72}$	0.366 ± 0.008	3.88 ± 0.84
Best fit	0.140	1.59	0.364	4.07
$r > 10h^{-1}$Mpc	0.143 ± 0.008	$1.57 \pm^{0.18\ 0.39\ 0.60}_{0.16\ 0.29\ 0.42}$	0.368 ± 0.006	3.98 ± 0.57
Best fit	0.145	1.55	0.369	3.88
Continuum/noise/LLS/DLA				
$r > 20h^{-1}$Mpc	0.172 ± 0.014	$1.38 \pm^{0.22\ 0.47\ 0.80}_{0.18\ 0.32\ 0.45}$	0.409 ± 0.007	3.50 ± 0.65
Best fit	0.176	1.33	0.410	3.50
$r > 10h^{-1}$Mpc	0.179 ± 0.008	$1.25 \pm^{0.11\ 0.23\ 0.38}_{0.10\ 0.18\ 0.26}$	0.402 ± 0.005	3.63 ± 0.45
Best fit	0.182	1.20	0.401	3.71
Continuum/noise/forest metals				
$r > 20h^{-1}$Mpc	0.149 ± 0.016	$1.45 \pm^{0.32\ 0.70\ 1.26}_{0.24\ 0.42\ 0.58}$	0.367 ± 0.008	4.04 ± 0.80
Best fit	0.156	1.35	0.366	4.12
$r > 10h^{-1}$Mpc	0.148 ± 0.009	$1.50 \pm^{0.17\ 0.38\ 0.63}_{0.15\ 0.28\ 0.39}$	0.369 ± 0.006	4.01 ± 0.52
Best fit	0.149	1.47	0.369	4.07
Continuum/noise/LLS/ DLA/forest metals				
$r > 20h^{-1}$Mpc	0.193 ± 0.014	$1.13 \pm^{0.17\ 0.36\ 0.61}_{0.14\ 0.26\ 0.37}$	0.412 ± 0.007	3.66 ± 0.62
Best fit	0.197	1.09	0.412	3.58
$r > 10h^{-1}$Mpc	0.187 ± 0.007	$1.15 \pm^{0.10\ 0.20\ 0.31}_{0.09\ 0.17\ 0.25}$	0.404 ± 0.005	3.62 ± 0.43
Best fit	0.190	1.12	0.403	3.60

Table 1. Results of parameter fittings for the average of 30 synthetic datasets and noise matrix for a single noisy measurement. The input values are $b = 0.145$, $\beta = 1.58$, $b(1 + \beta) = 0.375$ and $\alpha = 3.8$. Errorbars are $1 - \sigma$ confidence limits except for β where we give the 1, 2 and 3 σ error bars.

Finally, we note the surprisingly high degree of degeneracy in the b-β plane. This degeneracy is present in all measurements of β (see e.g. [76]), but the high values of β of the Lyman-α forest compared to typical galaxy populations makes it particularly acute. At $\beta = 1.5$, the power spectrum of purely radial modes has $(1 + 1.5)^2 = 6.25$ times more power than purely transverse modes and hence is measured much more precisely. If one measures just radial modes, the data would exhibit a perfect $b(1 + \beta)$ degeneracy, which is only relatively weakly broken by the measurements of the low μ modes.

5 Results with the observed data

In this section we discuss results derived from the observed data. Figure 11 illustrates, qualitatively, our 3-d measurement of structure traced by the Lyman-α forest, and it gives an idea of the relevant scales. The inset panel shows the distribution of BOSS quasars in a patch of sky $140' \times 70'$, corresponding to $170\,h^{-1}$ Mpc $\times\, 85\,h^{-1}$ Mpc (comoving) at $z = 2.5$ for our adopted cosmological model. The main panel plots the spectra of those quasars marked

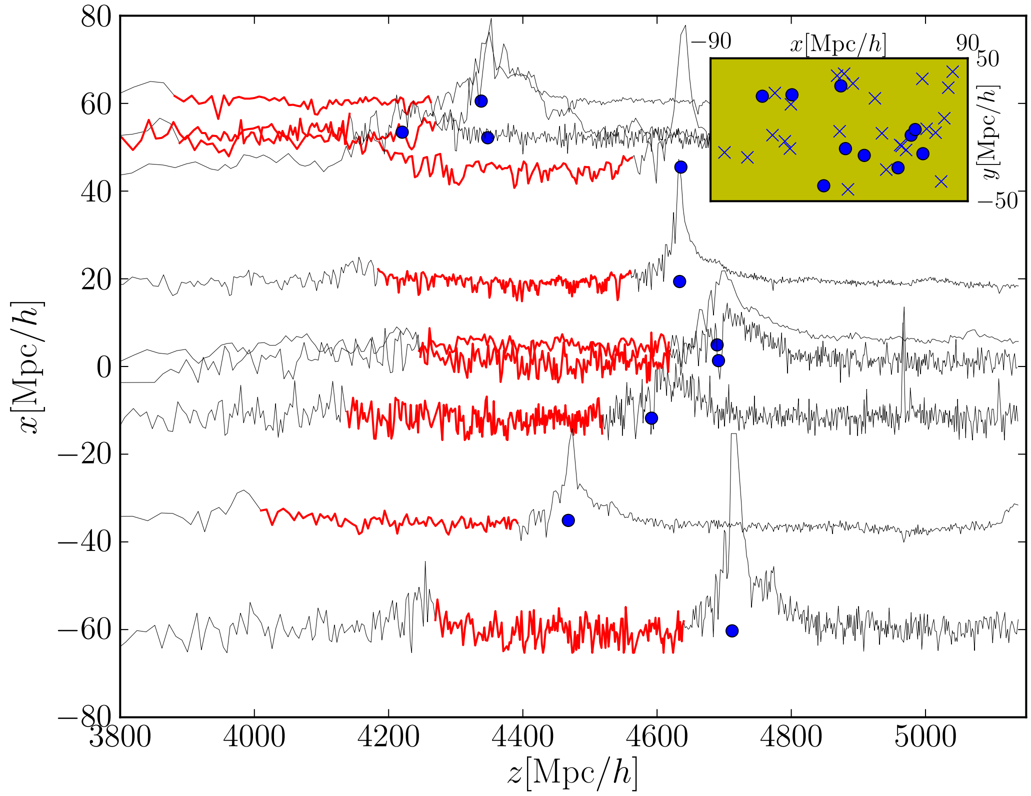

Figure 11. Illustration of 3-d Lyman-α forest measurement. Inset panel shows the distribution of BOSS quasars in a patch of sky $170\,h^{-1}$ Mpc × $85\,h^{-1}$ Mpc (comoving size at $z = 2.5$). The main panel plots spectra of the ten quasars marked by filled circles in the inset. Other quasars are included in the analysis but omitted here for clarity. Wavelengths have been converted to equivalent comoving distance for Lyman-α, and the Lyman-α forest regions are highlighted in red. Note that the scales for the vertical and horizontal axes are much different – in fact the lines of sight are much closer together, relative to their length, than they appear here.

in the inset panel by circles; we have omitted other quasars in the field (shown by ×'s in the inset) to preserve clarity. In the main panel, observed wavelengths have been converted to redshifts for the Lyman-α transition and from redshift to comoving distance using the standard cosmology. Note the radical difference in scale of the two axes in the main panel: the Lyman-α forest of a typical spectrum (highlighted in red) spans $\sim 400\,h^{-1}$ Mpc, while typical transverse separations of neighbouring quasars are a few h^{-1} Mpc up to $\sim 20\,h^{-1}$ Mpc.

Next we illustrate our continuum fitting process in Figure 12 for the observed data and the synthetic data. The left-hand panel shows the mean continuum, while the right hand panel shows the mean absorption *up to an arbitrary scaling factor* (since it is completely degenerate with unabsorbed continuum level). It is important to stress that no information from the actual measured data went into this first generation of synthetic data. The continua in the synthetic data were created from the *low-redshift* Hubble Space Telescope observations and the agreement between real and synthetic data attests to the universality of the quasar

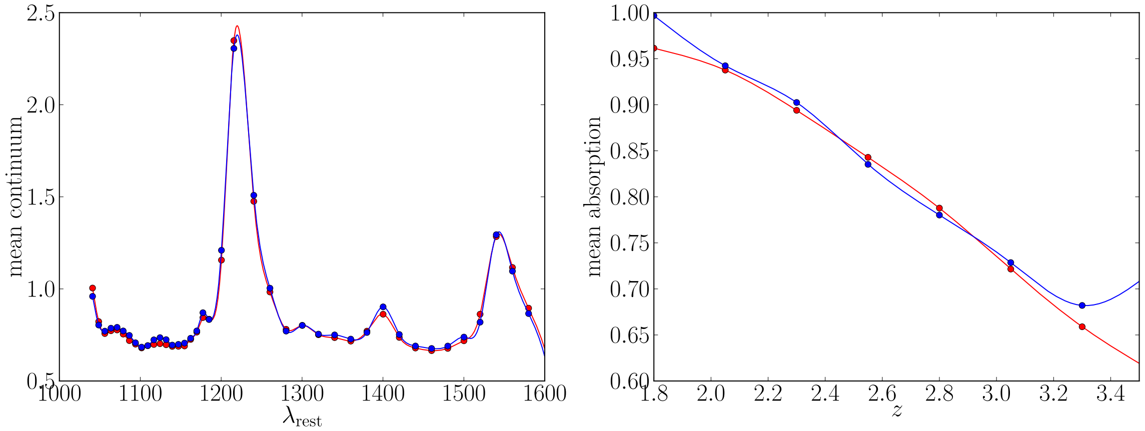

Figure 12. The mean continuum fits (left) and mean absorption fits (right), in arbitrary units, after thirty iterations of the continuum fitting pipeline for both the observed data (blue) and synthetic data (red). The curves diverge for $z < 2$ and $z > 3$ as there are virtually no data in those regions (see right panel of Figure 3).

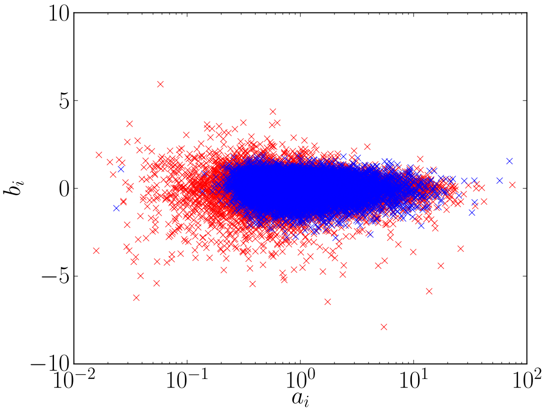

Figure 13. Distribution of parameters a_i and b_i for the data (red) and synthetic data (blue).

mean continuum. We do not plot the measurement errors here, since the errors on the final correlation function are derived using the measured correlations in the data, and these include all the extra variance due to continuum errors.

We illustrate the distribution of parameters a_i and b_i for real and synthetic data in Figure 13. The distribution for the observed data is considerably wider than that for the synthetic data. This is most likely due to the presence of the spectro-photometric errors in the data, but one should not exclude, in principle, a wider intrinsic variation in the shape of the quasar continua. However, we have shown that fitting out these parameters does not affect the derived correlation function.

In Figure 14 we plot the variance per pixel of δ_F, after the observational noise has been subtracted. To translate this quantity into a theory observable, one must take into account the effects of the pixel size, the point spread function of the telescope, and the variance

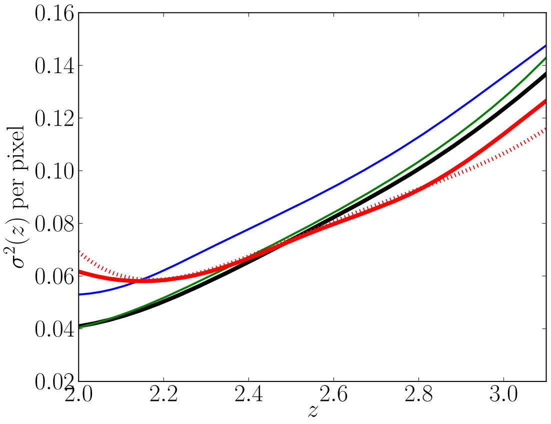

Figure 14. The remaining variance per pixel in the forest as a function of redshift after the observational noise has been subtracted. The thick black line is the value of the default synthetic data, while the blue line is the synthetic data with LLS/DLA added and the green line is with the forest metals added. The thick red line is for the data, and the dotted red line is for an alternative continuum fit to the data [77].

introduced by the continuum fitting. We see that at relatively high redshift the theoretical expectation obtained from the synthetic data sets agrees within $\sim 10\%$, although there is an unexpected upward trend in the residual variance at redshifts below 2.4. This will be further discussed in §7. The solid red curve describes the data with our standard continuum determination, while the dotted curve shows the results of using an alternative continuum fitting method that will be described elsewhere [77].

Finally we proceed to the actual results on the correlation function. Figure 15 displays the measured correlation function as a function of radius up to $100h^{-1}$Mpc, together with the best fit theory, after averaging over angle (upper plot), and also on a two-dimensional grid of transverse ($r_\perp$) and radial ($r_\|$) separation (two lower plots). These plots demonstrate the main two results of this paper: we have detected correlations in Lyman-α absorption in our 3-dimensional measurement out to comoving separations of $60\,h^{-1}$ Mpc, and we have detected strong redshift-space anisotropy in the flux correlation function. Both results are established at high statistical significance, and our measurements are consistent with the predictions of a standard ΛCDM cosmological large-scale structure model augmented by a well motivated 3-parameter description of the relation between the Lyman-α forest and the underlying dark matter distribution. The parameter values — describing linear bias, linear redshift-space distortion, and redshift evolution — are in accord with *a priori* theoretical expectations, once we account for the impact of high column density absorbers and metal lines.

We note also that our parameter errors are substantially larger for measurements on the real data relative to the mocks (not counting β, where the errors are highly dependent on the central value). Because the errors depend entirely on the measured correlation function in each case, this implies a substantially different correlation function between the two cases. We have not investigated carefully, but we suspect the difference is related to the large variance at low z in the real data (Figure 14), which is apparently not entirely compensated by boosted large-scale signal.

In Figure 16 we show the actual measured data-points of $\xi_F(r, \mu, z)$ in 30 panels for each bin in μ and z, together with the best-fit theory to guide the eye. We plot the square root of the diagonal elements of the covariance matrix as error bars, but measurements are correlated and therefore one should not attempt to evaluate χ^2 by eye. The points in this figure are our actual measurements used in the fitting of the bias parameters. For easier visualization of the results, we also convert them to a few alternative forms. In Figure 17 we average over redshift bins and radial bins in some cases and plot the same datapoints as a function of μ.

In Figure 18 we convert the ten μ measurements, averaged over redshift, into measurements of the multipoles of the correlation function. We perform the same operation for the best-fit theory. Our results are in striking agreement with the predictions of the standard linear (i.e., extended Kaiser) theory: we see a strong monopole correlation function that is proportional to the linear theory matter correlation function and a strong detection of a negative quadrupole, which is the signature of linear theory large-scale redshift-space distortions. (Note that the mean removal effect and the uneven redshift of individual points create a small $\ell = 6$ moment for the theoretical predictions, even though the $\ell = 6$ moment is exactly zero in pure linear theory.)

These results confirm that we have detected correlations in the Lyman-α forest flux in three dimensions out to a much larger scale than in previous measurements, and that we have detected the linear redshift distortions for the first time in the Lyman-α forest. This demonstrates that, on the large scales in which these linear correlations are measured, the dominant source of the Lyman-α forest transmission variations arise from gravitational instability of primordial mass fluctuations.

To put this claim on a more quantitative basis, we fit the bias parameters as described in section 4.5. The results are plotted in Figure 19 and given in Table 2. The best-fit χ^2 values are 233 with 237 degrees of freedom when fitting with points $r > 20h^{-1}$Mpc and 281 with 297 degrees of freedom when using points with $r > 10h^{-1}$Mpc.

Compared to the simulations of reference [39], the data prefer lower values of β and higher values of bias, although the product $b(1 + \beta)$ is of the right magnitude. We also see a somewhat lower evolution with redshift. The formal probability that $\beta > 1.5$ is $\sim 3 \times 10^{-3}$ ($\sim 1 \times 10^{-3}$ for $r > 10h^{-1}$Mpc fitting; both cases probably dominated by the MCMC sampling noise), and $\beta > 1.0$ about 10% (24% for $r > 10h^{-1}$Mpc fitting). For fitting each redshift bin individually, the probabilities are 11%, 24% and 48% for $\beta > 1.5$ and 33%, 48% and 65% for $\beta > 1.0$ for the lowest, medium and highest redshift bin respectively. We see that the low value β for a single β fit is driven a lot by the lowest redshift bin. Our synthetic data sets show that metals and high column density systems (LLS/DLAs) can considerably lower the observed value of β. Clearly, more work will be required to explain in detail the values of bias parameters that we find.

To determine the distance to which we have formally detected correlations, we calculate the value of χ^2 for a model with no correlations and compare it to the χ^2 of the three parameter model. Using this criterion, we have detected correlations up to a distances $60h^{-1}\text{Mpc} < r < 100h^{-1}$Mpc at 3-$\sigma$ ($\Delta\chi^2 > 9$), and up to $70h^{-1}\text{Mpc} < r < 100h^{-1}$Mpc at 2-$\sigma$ ($\Delta\chi^2 > 4$). Similarly, we have detected redshift-space distortions with high significance. For $r > 20h^{-1}$Mpc, setting $\beta = 0$ while varying other parameters results in $\Delta\chi^2 \sim 48$, while using distance $r > 10h^{-1}$Mpc one gets $\Delta\chi^2 \sim 120$. These large $\Delta\chi^2$ values imply very high statistical confidence. We quote this numbers as "over 5σ" to conservatively take into account potential imperfections in the error matrix.

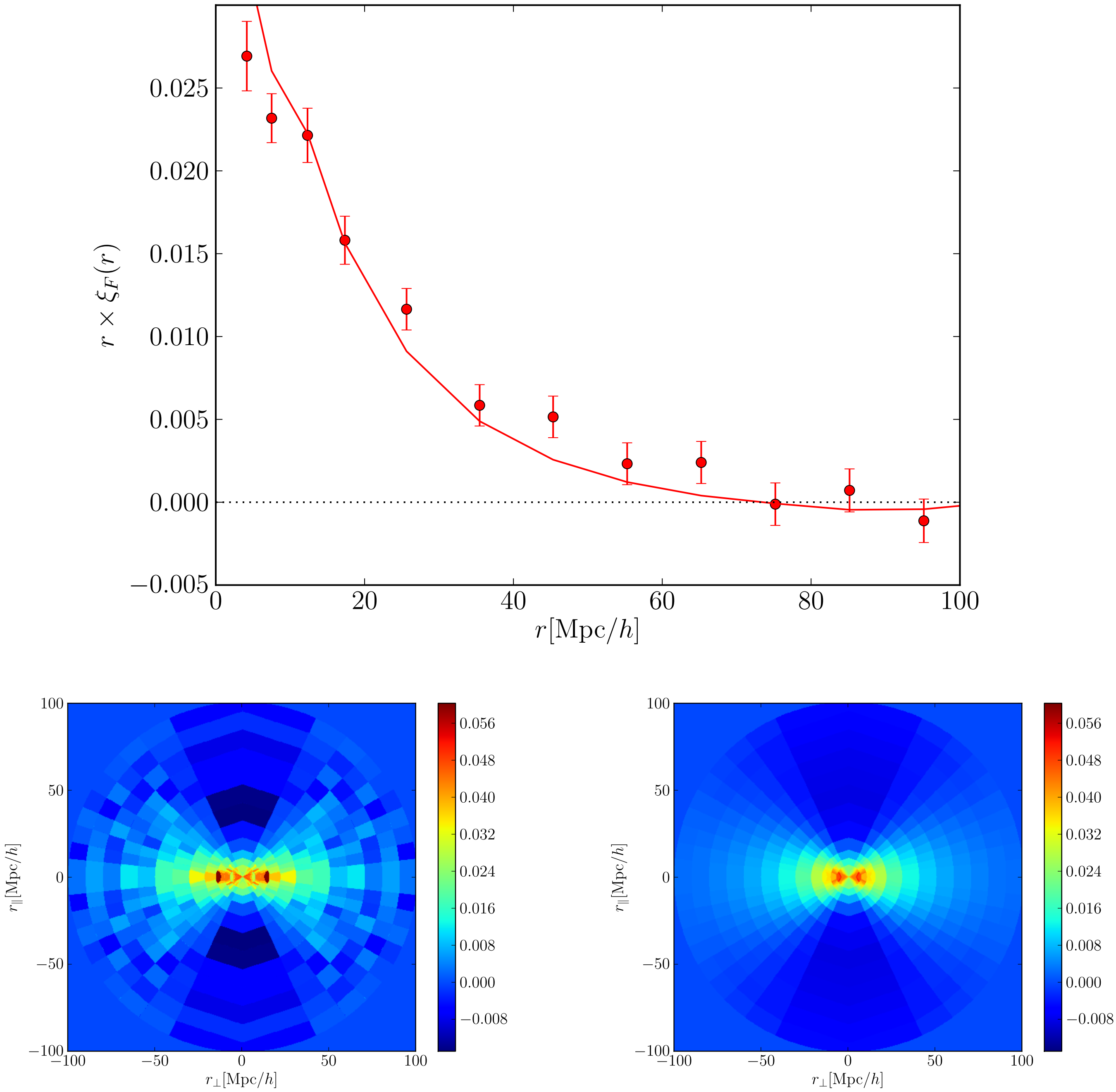

Figure 15. Primary measurement results in visual form. Top plot shows the monopole of the correlation function, together with a best-fit two-parameter $(b,\ \beta)$ linear model. The bottom two plots are redshift averaged data plotted in the plane $r_\perp - r_\parallel$, with each pixel plotted with the value corresponding to the nearest neighbor. The left panel corresponds to data, and the right panel to the corresponding best-fit theory.

Data	bias	β	$b(1+\beta)$	α
$r > 20h^{-1}$Mpc	0.197 ± 0.021	$0.71\pm^{0.21\ 0.49\ 0.87}_{0.16\ 0.27\ 0.39}$	0.336 ± 0.012	1.59 ± 1.55
$r > 10h^{-1}$Mpc	0.175 ± 0.012	$0.90\pm^{0.15\ 0.33\ 0.56}_{0.13\ 0.23\ 0.33}$	0.333 ± 0.008	2.09 ± 0.94
With LLS/DLA, $r > 20h^{-1}$Mpc	0.217 ± 0.025	$0.55\pm^{0.19\ 0.48\ 0.97}_{0.14\ 0.25\ 0.35}$	0.337 ± 0.014	2.99 ± 1.74
With LLS/DLA, $r > 10h^{-1}$Mpc	0.180 ± 0.013	$0.87\pm^{0.16\ 0.35\ 0.56}_{0.13\ 0.25\ 0.35}$	0.337 ± 0.009	3.11 ± 0.93
$r > 20h^{-1}$Mpc,				
1st z-bin $z = 2.0 - 2.2$	0.168 ± 0.033	$0.81\pm^{0.54\ 1.64\ 3.33}_{0.30\ 0.49\ 0.65}$	0.305 ± 0.017	/
2nd z-bin $z = 2.2 - 2.4$	0.167 ± 0.037	$0.97\pm^{0.82\ 2.55\ 3.86}_{0.39\ 0.62\ 0.78}$	0.330 ± 0.019	/
3rd z-bin $z = 2.4 - 3.0$	0.164 ± 0.047	$1.33\pm^{1.56\ 3.21\ 3.65}_{0.66\ 0.98\ 1.19}$	0.383 ± 0.033	/
$r > 20h^{-1}$Mpc,				
$\mu > 0.1$	0.200 ± 0.021	$0.63\pm^{0.19\ 0.45\ 0.78}_{0.15\ 0.27\ 0.38}$	0.325 ± 0.013	1.57 ± 1.77
$\mu < 0.9$	0.206 ± 0.023	$0.65\pm^{0.21\ 0.48\ 0.80}_{0.16\ 0.28\ 0.38}$	0.341 ± 0.013	1.33 ± 1.60
$r > 20h^{-1}$Mpc,				
1041Å$< \lambda_{\rm rest} <$ 1120Å	0.207 ± 0.050	$0.36\pm^{0.30\ 0.96\ 2.66}_{0.18\ 0.30\ 0.35}$	0.285 ± 0.054	-9.31 ± 13.87
1120Å$< \lambda_{\rm rest} <$ 1185Å	0.118 ± 0.032	$1.62\pm^{1.54\ 2.98\ 3.35}_{0.74\ 1.08\ 1.31}$	0.311 ± 0.024	3.40 ± 2.91
$r > 20h^{-1}$Mpc,				
$g < 20.5$	0.137 ± 0.035	$1.10\pm^{1.01\ 2.94\ 3.82}_{0.45\ 0.71\ 0.90}$	0.295 ± 0.042	-9.10 ± 7.55
$g > 20.5$	0.162 ± 0.035	$1.20\pm^{0.99\ 2.78\ 3.72}_{0.45\ 0.71\ 0.89}$	0.357 ± 0.022	2.64 ± 2.36

Table 2. This table shows the results of the parameter fittings for the data and various systematics checks. All error bars are $1-\sigma$ error bars, except for the β parameter in which case we give 1,2 and $3-\sigma$ confidence limits.

6 Systematic effects, cross-checks and data-splits

In the previous section we have shown results of fitting the bias parameters to the measured correlation function. In this section we investigate how stable these results are against various divisions of the data. We always perform two basic cuts on the data: we never use pairs of pixels from the same quasar, and we never use pairs of pixels that are separated by less than 1.5 Å. In addition, we have performed several tests by splitting the data in a variety of ways. Generally, for each split we measure the bias parameters and compare these to the full dataset. If the combined data were inconsistent with a subset of the data, this might indicate some unknown systematic error. For example, if what we are measuring are truly cosmological fluctuations, our results should be the same regardless of using bright or faint quasars. We perform the following data splits:

- *Perform the fitting excluding the lowest μ or the highest μ bin.* In both cases, the central values shift and error bars increase, but the overall fit is not driven by either low and high μ values alone.

- *Perform the fitting as a function of separation. We fit for bias-β parameters in 5 radial bins.* Results of this test are plotted in Figure 21. The value of α was fixed at the fiducial value of $\alpha = 3.9$, but α is not degenerate with other parameters.

- *Perform the fitting as a function of redshift.* We fit for the bias and β parameters independently in each redshift bin. Results of this test are plotted in Figure 20.

- *Split by Lyman-α forest range.* We divided the total data into two segments 1041Å $-$ 1120Å and 1120Å $-$ 1185Å. Note that each of these sets contain only a quarter of the original pairs (the remaining half being in the cross-correlations between the two halves). The measurement of α is very poor using the short wavelength end because there is a dearth of pixels at relatively high redshift in this case.

- *Split by quasar brightness.* We divided by g-band magnitude into quasars brighter or fainter than 20.5. Note that each of these sets contain only a quarter of the original pairs. The measurement of α is very poor using the bright quasars because there is a dearth of pixels at relatively high redshift in this case.

None of these tests resulted in a detection of a significant systematic effect. Results of this exercise are also presented in Table 2. Note, however, that these splits are not entirely satisfactory, as the differences are hard to test at precision approaching the overall precision. E.g., in the case where μ is restricted to be > 0.1, $b(1+\beta)$ changes from 0.336 ± 0.012 to 0.325 ± 0.013. This seems like a small change, but on the other hand it is almost 1 σ when only a small fraction of the data is removed. If we think of $\mu > 0.1$ and $\mu < 0.1$ as two independent measurements of $b(1+\beta)$, the implied difference between the values is 0.074 ± 0.034, i.e., a 2.2 σ difference. There is no compelling reason to believe this is evidence for a systematic error (this was chosen out of several possibilities as an example of a big difference), but it highlights the fact that a split can differ by ~ 6 σ by the scale of the full-sample errors (i.e., 0.074/0.012) yet still not be decisive evidence for a systematic problem, i.e., big effects could hide in this kind of test. In the cases where the sample of quasars is split, the errors often expand catastrophically due to a reduction in cross-correlation pairs, so in the end the tests are not actually very powerful. In the future, we may try to squeeze out more powerful tests by, e.g., fitting for some simple parameterized dependence of b and β on quasar magnitude (or other properties), continuing to use the full data set in the fit.

7 Discussion

We measure the redshift-space distortion parameter β to be between 0.44 and 1.2 at the 95% confidence level. This value is lower than the theoretical prediction from numerical simulations of the Lyman-α forest in [39], $\beta \simeq 1.47$, with small error bars for the particular model that was analyzed there. We have shown here that forest metal contamination and LLS/DLAs may help explain this discrepancy: in the simple model we have adopted to include LLS/DLAs in our synthetic data, these systems alone lower the fitted value of β from 1.58 to $\beta \simeq 1.33$, while forest metals alone lower it to $\beta \simeq 1.35$. The two effects together can lower it to $\beta \simeq 1.09$ (we used $\beta = 1.58$ to start here, instead of 1.47, because we had not yet adjusted the prediction of [39] to reflect a modern cosmological model). Note that we have removed by-eye-identified DLAs from the data set, so any effect must be coming from ones that are missed, due to low column density and/or noise in the data.

When we include identified DLAs in the analysis (Table 2), b and β change in the direction predicted by the mocks, but it is actually quite difficult to estimate whether this change is consistent with a small or large fraction of the systems relevant to b and β being included in this identified sample. Since optical depth is additive, the observed flux fluctuations are $\delta_f(x) = (1+\delta_F(x))(1+\delta_D(x))$ where $\delta_F(x)$ is low density forest and $\delta_D(x)$ high density. At this point one might be tempted to approximate this as $\delta_f(x) \simeq \delta_F(x) + \delta_D(x)$, which makes it straightforward to estimate b and β for one of the fields given the other two. This would make it possible to interpret the mocks as predicting b and β for the DLA/LLSs, and then estimate exactly how, given that model, the observations with DLA/LLSs should be related to ones without them. Unfortunately, as shown by [78, 79], this linearized calculation is mathematically nonsensical. A careful study of a term like $\langle \delta_F(x)\delta_D(x)\delta_F(y)\rangle$ shows that this is *not* generically smaller than $\langle \delta_D(x)\delta_F(y)\rangle$ – a naive Taylor series approach is not valid because the product $\delta_F(x)\delta_D(x)$ applies locally, i.e., at a small-scale point where fluctuations are not small. Pursuing this calculation to next-to-leading order shows that gravitational evolution generically leads to a modulation of this local product by large scale density modes, so the composite field $\delta_F(x)\delta_D(x)$ appears as a standard linearly biased field on large scales.

Somewhat puzzlingly, adding real DLAs increased our α result substantially (Table 2), while the mocks actually predict a reduction. Clearly there are some imperfections in our treatment, although sub-DLAs which are included in the mocks but not identified in real data could in *principle* account for these oddities. We emphasize here that the degree to which LLS/DLA and metal lines may lower the value of β ought to depend on the way in which these systems are inserted in our synthetic data sets, and on the way they are cross-correlated with the Lyman-α forest. The model we have adopted to insert the DLA/LLS systems in our mock spectra should only be considered as an illustrative example of their plausible effect. An observationally calibrated physical model of the distribution of these systems will be required before reliable predictions can be made of their impact on the value of β.

As discussed in the introduction, simulations with lower resolution/smoothing scale than [39] (including [39]'s own low resolution simulations) find considerably lower value of $\beta \sim 1.0$ ([40], Martin White private communication), so it is possible that a low β that survives other explanations is an indication of smoother-than-expected small-scale gas, or a flaw in the HPM modeling of pressure used in [39].

The second parameter that we measure is the bias. Bias is, of course, completely degenerate with the assumed value of σ_8, which at the moment is known to about 3% in the simplest LCDM model [80]. The parameter that we are really measuring from the Lyman-α forest observations is the product $b\sigma_8(z \sim 2.25)$. We assumed a value of $\sigma_8(z = 0) = 0.8$ and our inferred bias varies as the inverse of $\sigma_8(z = 2.25)$ (where we mean σ_8 loosely – really, the linear power on the scale that we are measuring). Our bias is constrained to be between 0.16 and 0.24, a value which is considerably higher than the $b \sim 0.13$ obtained in [39] and [41] (the latter verified by later simulations of Martin White with higher resolution). [40] did obtain a higher value $b \sim 0.18$ from very low resolution simulations, but the numerical smoothing in these simulations is almost certainly much larger than any physical smoothing of the IGM. These theoretical numbers are for the uncontaminated forest; metal contamination associated with forest absorption negligibly affects bias, but LLS/DLAs can raise the bias considerably, by some 20% in synthetic data. The effect of including or excluding quasars marked as DLAs is about 10% on the bias in our data. This is consistent with the fact that our non-DLA flagged sample is still likely to be contaminated with high column density systems at some level (see §2). Regardless of the cause, the higher-than-expected bias may improve BAO detection by creating a higher-than-forecast signal.

There is some evidence that our lowest redshift bin is the most problematic. From Figure 20 it is clear that it contributes a lot to the overall signal on β and that it drives the low value of β observed here. Moreover, Figure 14 indicates a greater variance than expected at the lowest values of z (this expectation is an extrapolation of the trend in the SDSS observational measurement of [9], not necessarily a simulation prediction). The cause of this enhanced variance is not yet understood. The bias evolution parameter α is similarly smaller than expected, indicating extra power at low z on large scales, although this is only significant enough to be suggestive. At current sensitivity, the data are consistent with a constant β and bias, which suggests that any contaminant that is affecting the large-scale correlation is itself a tracer of the large scale structure.

Irrespective of these uncertainties on the values of b and β and their implications for the physics of the Lyman-α forest, which will need to be further investigated in the future, one main conclusion stands out from this work: the correlation function of the Lyman-α forest on the scales of 10 to $60h^{-1}$Mpc bears the signature of redshift distortions consistently with the growth of linear density perturbations by gravitational instability. The detailed physics of the Lyman-α forest may still be influenced by other processes on smaller scales, such as galactic winds and outflows from quasar jets, but on the large scales examined here, linear gravitational evolution must be the principal process at work.

What are the main improvements that are desirable for the next iteration of the analysis of the BOSS Lyman-α forest data? We have reasons to believe that our analysis is sufficient for the goals in this paper, mainly because of the good fit we obtain to the standard theoretical model and the tests performed in §6. However, the next iteration may require a more sophisticated analysis, especially to better understand the lowest redshift bin. First, we need to remove the potential sources of systematics by subtracting the 'signal' measured red-ward of the Lyman-α emission line. This should eliminate potential contamination from low-redshift metal lines as well as any systematics associated with, e.g., imperfect sky subtraction. With the present quasar sample we have performed this test; however the sample used here did not include low-z quasars that would be required to do this subtraction in the lowest redshift bin. The two higher redshift bins did not show any measurable deviation from zero. Note, however, that such methods will not solve the problem of metals lines that arise exclusively (or almost exclusively) in the forest, as described in § 3.3. Second, the continuum fitting could perhaps be improved by going beyond a fixed continuum model (e.g. [77]) and more thoroughly investigating the spectro-photometric errors in the data. Third, our results have shown that there is a real need to better understand and filter out the high column density and metal-line systems systems present in the data. These should be identified, possibly using the absorption features red-ward of the Lyman-α emission line, and either corrected or removed from the data. Only the approximate impact of forest metals has been explored here. As the survey grows, measurements of metal clustering and scatter in absorption strength will be included in the analysis. This will also provide greater precision and, perhaps, sensitivity to weaker metal lines. We have not explored in this paper the impact of metals associated with LLS/DLAs, and this is something we intend to address. Fourth, additional physical effects might complicate the biasing of the observed field with respect to the dark matter, such as temperature and ionization fluctuations in the intergalactic medium [44] (although see [81] for a possible way to constrain these from the 1-dimensional Lyman-α forest). These do not affect the measurement *per se*, but they do affect its interpretation. Finally, the method for calculating the correlation function (or power spectrum) should be made more optimal. This should include a better treatment of the evolution of the flux across the forest and an appropriate form of inverse variance weighting. The full problem is computationally intractable, but one could apply the inverse covariance weighting on a per quasar basis, an approximately optimal weighting suggested in [55, 56], or do the full problem on coarse pixels.

One of the main ultimate goals of the measurement of the Lyman-α forest correlation function is to infer the angular diameter distance $D_A(z)$ and the Hubble constant $H(z)$ at the observed redshifts, using the position of the BAO peak. However, even at the smaller scales at which we have made our measurements here, cosmological constraints might be obtainable on the product $D_A(z)H(z)$ through the Alcock-Paczyński method [82–84]. To illustrate the statistical potential of this test, we have attempted to fit the observed data assuming an Einstein-de-Sitter cosmology. If we simply rescale the radial and transverse distances, keeping a constant form for the linear theory power spectrum, spurious higher multipoles appear in the redshift-space correlation function. This results in the best fit χ^2 degrading by $\Delta\chi^2 \sim 10$ when using only $r > 20h^{-1}$Mpc, and by $\Delta\chi^2 \sim 21$ when using $r > 10h^{-1}$Mpc. This procedure is not a fully geometrical form of the Alcock-Paczyński test, since we have assumed that the real space correlation function has the shape predicted by ΛCDM, but it shows that our data already have enough power to detect any large deviations from the spacetime metric of a flat, Λ-dominated universe.

A much stronger change appears if, in addition to changing distances, we also change the underlying theory. The CDM correlation function for $\Omega_m = 1$ passes through zero in the radial direction at $\sim 28h^{-1}$Mpc, clearly at odds with our data (see e.g. Figure 15). The best-fit χ^2 in this case increases by $\Delta\chi^2 \sim 48$ when fitting $r > 20h^{-1}$Mpc and by $\Delta\chi^2 \sim 88$ when fitting with $r > 10h^{-1}$Mpc. We caution that this procedure has not been tested with synthetic data; fitting cosmological parameters goes beyond the scope of this paper. This shows, however, that the shape of the correlation function may contain substantial cosmological information in addition to the BAO feature if systematic errors can be well controlled.

8 Conclusions and Prospects

For more than a decade, 1-dimensional analyses of the Lyman-α forest have provided a powerful quantitative tool for probing structure in the high-redshift universe. The BOSS quasar sample makes it possible, for the first time, to treat large-scale structure in the Lyman-α forest as a truly 3-dimensional phenomenon. Although this first-year BOSS quasar sample is only 10% of the anticipated final sample, it is already several times larger than the largest previous sample used for cosmological analysis of the Lyman-α forest [9]. It is similar in size to the entire sample of $z > 2.1$ quasars from SDSS-I and SDSS-II [58], and the order-of-magnitude higher surface density of BOSS quasars makes it a much more powerful sample for 3-dimensional measurements. We have achieved high-significance detection of the angle-averaged flux correlation function out to comoving separation of $60\,h^{-1}$ Mpc, and the shape of this correlation function agrees well with the predictions of a standard ΛCDM cosmological model.

Our measurements show the clear signature of redshift-space anisotropy induced by large-scale peculiar velocities. The agreement of the observed anisotropy with the linear theory prediction of the extended Kaiser model confirms the standard model of the Lyman-α forest as structure that originates in the gravitational instability of primordial density fluctuations [30–33, 85].

We have fit our measurements with a 3-parameter model that describes the linear bias of the forest (b), the redshift-space distortion (β), and the redshift-evolution of the correlation amplitude (α). Our estimated parameter values are within the range of theoretical predictions, though the value of β appears somewhat low (see §7). Our synthetic data tests suggest that this low β may be a consequence of high column density systems (LLS/DLA) and metal-line absorption within the forest. Statistical errors estimated internally from the data agree well with external estimates based on the synthetic data sets, which suggests that we have identified any observational or physical effects that have a large impact on our measurements.

The tests in §7 show that assuming either an $\Omega_m = 1$ spacetime metric or an $\Omega_m = 1$ CDM matter power spectrum leads to substantially worse agreement with our measurements. However, we have not attempted to derive cosmological parameter constraints, instead fitting values of b, β, and α assuming an underlying ΛCDM cosmology. Previous studies using the 1-dimensional flux power spectrum have inferred the slope and amplitude of the matter power spectrum by using cosmological simulations to predict the bias of the flux power spectrum (including its scale dependence) from first principles. Even after marginalizing over uncertainties in the IGM equation of state, these studies yield valuable cosmological constraints (e.g., [4–7, 18, 20]

The BOSS Lyman-α forest measurements will allow these tests to become much more powerful. The measurement of the 1-dimensional power spectrum will itself become much more precise with the large BOSS data sample, and division of the data set into many subsamples of redshift, data quality, and quasar properties will allow careful cross-checks for systematic errors. Much stronger constraints can be obtained from three-dimensional measurements, because of the additional information contained in the cross-correlation of parallel lines of sight and because they allow for strong tests based on redshift space distortions and the cosmological dependence of the angular diameter distance and expansion rate. Fully exploiting these data will require considerable analysis of the systematic effects of the DLA/LLS and metal-line absorption and of additional physical effects on the correlation function, such as those due to variations in the ionizing background or in the temperature-density relation induced by helium reionization [44]. However, BOSS data will provide many measurements with which to constrain these models and test for observational or theoretical systematics.

The design goal of the BOSS quasar survey is to measure the angular diameter distance $D_A(z)$ and Hubble parameter $H(z)$ at $z \approx 2.5$ from the BAO feature in Lyman-α forest clustering [24]. Forecasts using the formalism of [55] imply 1σ constraints of 7.7% and 3.0% on these two quantities, respectively, from the full survey. These errors are strongly correlated (similar to the b-β degeneracy found in this paper), so it is more meaningful to quote the forecast error on an overall distance scale dilation factor, which is 1.9%. Our present measurement of clustering on sub-BAO scales is based on 10% of the full BOSS data sample and on first-pass versions of the spectroscopic reduction pipeline and Lyman-α forest analysis procedures. The good agreement that we find with theoretical expectations reinforces the promise of the Lyman-α forest as a tool to map the high-redshift universe, to measure its expansion via BAO, and to thereby constrain the origin of cosmic acceleration.

Acknowledgements

We thank Martin White, David Kirkby and many others for useful discussions.

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

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

This work was supported in part by the U.S. Department of Energy under Contract No. DE-AC02-98CH10886. AF and JM are supported by Spanish grant AYA2009-09745. Some computations were performed on CITA's Sunnyvale clusters which are funded by the Canada Foundation for Innovation, the Ontario Innovation Trust, and the Ontario Research Fund.

References

[1] R. Lynds, *The Absorption-Line Spectrum of 4c 05.34*, *ApJL* **164** (Mar., 1971) L73+.

[2] J. E. Gunn and B. A. Peterson, *On the Density of Neutral Hydrogen in Intergalactic Space.*, *ApJ* **142** (Nov., 1965) 1633–1641.

[3] M. Rauch, *The Lyman Alpha Forest in the Spectra of QSOs*, *ARAA* **36** (1998) 267–316, [astro-ph/].

[4] R. A. C. Croft, D. H. Weinberg, M. Pettini, L. Hernquist, and N. Katz, *The Power Spectrum of Mass Fluctuations Measured from the LYalpha Forest at Redshift Z=2.5*, *ApJ* **520** (July, 1999) 1–23.

[5] P. McDonald, J. Miralda-Escudé, M. Rauch, W. L. W. Sargent, T. A. Barlow, R. Cen, and J. P. Ostriker, *The Observed Probability Distribution Function, Power Spectrum, and Correlation Function of the Transmitted Flux in the Lyα Forest*, *ApJ* **543** (Nov., 2000) 1–23.

[6] R. A. C. Croft, D. H. Weinberg, M. Bolte, S. Burles, L. Hernquist, N. Katz, D. Kirkman, and D. Tytler, *Toward a Precise Measurement of Matter Clustering: Lyα Forest Data at Redshifts 2-4*, *ApJ* **581** (Dec., 2002) 20–52.

[7] M. Viel, M. G. Haehnelt, and V. Springel, *Inferring the dark matter power spectrum from the Lyman α forest in high-resolution QSO absorption spectra*, *MNRAS* **354** (Nov., 2004) 684–694.

[8] P. McDonald, U. Seljak, R. Cen, D. Shih, D. H. Weinberg, S. Burles, D. P. Schneider, D. J. Schlegel, N. A. Bahcall, J. W. Briggs, J. Brinkmann, M. Fukugita, Ž. Ivezić, S. Kent, and D. E. Vanden Berk, *The Linear Theory Power Spectrum from the Lyα Forest in the Sloan Digital Sky Survey*, *ApJ* **635** (Dec., 2005) 761–783.

[9] P. McDonald, U. Seljak, S. Burles, D. J. Schlegel, D. H. Weinberg, R. Cen, D. Shih, J. Schaye, D. P. Schneider, N. A. Bahcall, J. W. Briggs, J. Brinkmann, R. J. Brunner, M. Fukugita, J. E. Gunn, Ž. Ivezić, S. Kent, R. H. Lupton, and D. E. Vanden Berk, *The Lyα Forest Power Spectrum from the Sloan Digital Sky Survey*, *ApJS* **163** (Mar., 2006) 80–109.

[10] D. G. York *et. al.*, *The Sloan Digital Sky Survey: Technical Summary*, *AJ* **120** (Sept., 2000) 1579–1587.

[11] C. Stoughton, R. H. Lupton, M. Bernardi, M. R. Blanton, S. Burles, F. J. Castander, A. J. Connolly, D. J. Eisenstein, J. A. Frieman, G. S. Hennessy, R. B. Hindsley, Ž. Ivezić, S. Kent, P. Z. Kunszt, B. C. Lee, A. Meiksin, J. A. Munn, H. J. Newberg, R. C. Nichol, T. Nicinski, J. R. Pier, G. T. Richards, M. W. Richmond, D. J. Schlegel, J. A. Smith, M. A. Strauss, M. SubbaRao, A. S. Szalay, A. R. Thakar, D. L. Tucker, D. E. Vanden Berk, B. Yanny, and the SDSS collaboration, *Sloan Digital Sky Survey: Early Data Release*, *AJ* **123** (Jan., 2002) 485–548.

[12] J. A. Smith, D. L. Tucker, S. Kent, M. W. Richmond, M. Fukugita, T. Ichikawa, S.-i. Ichikawa, A. M. Jorgensen, A. Uomoto, J. E. Gunn, M. Hamabe, M. Watanabe, A. Tolea, A. Henden, J. Annis, J. R. Pier, T. A. McKay, J. Brinkmann, B. Chen, J. Holtzman, K. Shimasaku, and D. G. York, *The u'g'r'i'z' Standard-Star System*, *AJ* **123** (Apr., 2002) 2121–2144, [astro-ph/0201143].

[13] J. R. Pier, J. A. Munn, R. B. Hindsley, G. S. Hennessy, S. M. Kent, R. H. Lupton, and Ž. Ivezić, *Astrometric Calibration of the Sloan Digital Sky Survey*, *AJ* **125** (Mar., 2003) 1559–1579, [astro-ph/0211375].

[14] N. Padmanabhan, D. J. Schlegel, D. P. Finkbeiner, J. C. Barentine, M. R. Blanton, H. J. Brewington, J. E. Gunn, M. Harvanek, D. W. Hogg, Ž. Ivezić, D. Johnston, S. M. Kent, S. J. Kleinman, G. R. Knapp, J. Krzesinski, D. Long, E. H. Neilsen, Jr., A. Nitta, C. Loomis, R. H. Lupton, S. Roweis, S. A. Snedden, M. A. Strauss, and D. L. Tucker, *An Improved Photometric Calibration of the Sloan Digital Sky Survey Imaging Data*, *ApJ* **674** (Feb., 2008) 1217–1233, [astro-ph/0703454].

[15] J. E. Gunn, W. A. Siegmund, E. J. Mannery, R. E. Owen, C. L. Hull, R. F. Leger, L. N. Carey, G. R. Knapp, D. G. York, W. N. Boroski, S. M. Kent, R. H. Lupton, C. M. Rockosi, M. L. Evans, P. Waddell, J. E. Anderson, J. Annis, J. C. Barentine, L. M. Bartoszek, S. Bastian, S. B. Bracker, H. J. Brewington, C. I. Briegel, J. Brinkmann, Y. J. Brown, M. A. Carr, P. C. Czarapata, C. C. Drennan, T. Dombeck, G. R. Federwitz, B. A. Gillespie, C. Gonzales, S. U. Hansen, M. Harvanek, J. Hayes, W. Jordan, E. Kinney, M. Klaene, S. J. Kleinman, R. G. Kron, J. Kresinski, G. Lee, S. Limmongkol, C. W. Lindenmeyer, D. C. Long, C. L. Loomis, P. M. McGehee, P. M. Mantsch, E. H. Neilsen, Jr., R. M. Neswold, P. R. Newman, A. Nitta, J. J. Peoples, J. R. Pier, P. S. Prieto, A. Prosapio, C. Rivetta, D. P. Schneider, S. Snedden, and S.-i. Wang, *The 2.5 m Telescope of the Sloan Digital Sky Survey*, *AJ* **131** (Apr., 2006) 2332–2359, [astro-ph/0602326].

[16] J. E. Gunn, M. Carr, C. Rockosi, M. Sekiguchi, K. Berry, B. Elms, E. de Haas, Ž. Ivezić, G. Knapp, R. Lupton, G. Pauls, R. Simcoe, R. Hirsch, D. Sanford, S. Wang, D. York, F. Harris, J. Annis, L. Bartozek, W. Boroski, J. Bakken, M. Haldeman, S. Kent, S. Holm, D. Holmgren, D. Petravick, A. Prosapio, R. Rechenmacher, M. Doi, M. Fukugita, K. Shimasaku, N. Okada, C. Hull, W. Siegmund, E. Mannery, M. Blouke, D. Heidtman, D. Schneider, R. Lucinio, and J. Brinkman, *The Sloan Digital Sky Survey Photometric Camera*, *AJ* **116** (Dec., 1998) 3040–3081, [astro-ph/9809085].

[17] M. Fukugita, T. Ichikawa, J. E. Gunn, M. Doi, K. Shimasaku, and D. P. Schneider, *The Sloan Digital Sky Survey Photometric System*, *AJ* **111** (Apr., 1996) 1748–+.

[18] U. Seljak, A. Makarov, P. McDonald, S. F. Anderson, N. A. Bahcall, J. Brinkmann, S. Burles, R. Cen, M. Doi, J. E. Gunn, Ž. Ivezić, S. Kent, J. Loveday, R. H. Lupton, J. A. Munn, R. C. Nichol, J. P. Ostriker, D. J. Schlegel, D. P. Schneider, M. Tegmark, D. E. Berk, D. H. Weinberg, and D. G. York, *Cosmological parameter analysis including SDSS Lyα forest and galaxy bias: Constraints on the primordial spectrum of fluctuations, neutrino mass, and dark energy*, *Phys. Rev. D* **71** (May, 2005) 103515–+.

[19] U. Seljak, A. Makarov, P. McDonald, and H. Trac, *Can Sterile Neutrinos Be the Dark Matter?*, *Physical Review Letters* **97** (Nov., 2006) 191303–+, [astro-ph/].

[20] U. Seljak, A. Slosar, and P. McDonald, *Cosmological parameters from combining the Lyman-α forest with CMB, galaxy clustering and SN constraints*, *Journal of Cosmology and Astro-Particle Physics* **10** (Oct., 2006) 14–+, [astro-ph/0604335].

[21] M. Viel and M. G. Haehnelt, *Cosmological and astrophysical parameters from the Sloan Digital Sky Survey flux power spectrum and hydrodynamical simulations of the Lyman α forest*, *MNRAS* **365** (Jan., 2006) 231–244.

[22] M. Viel, G. D. Becker, J. S. Bolton, M. G. Haehnelt, M. Rauch, and W. L. W. Sargent, *How*

Cold Is Cold Dark Matter? Small-Scales Constraints from the Flux Power Spectrum of the High-Redshift Lyman-α Forest, *Physical Review Letters* **100** (Feb., 2008) 041304–+, [arXiv:0709.0131].

[23] M. Viel, M. G. Haehnelt, and V. Springel, *The effect of neutrinos on the matter distribution as probed by the intergalactic medium*, *JCAP* **6** (June, 2010) 15–+, [arXiv:1003.2422].

[24] D. J. Eisenstein, D. H. Weinberg, E. Agol, H. Aihara, C. Allende Prieto, S. F. Anderson, J. A. Arns, E. Aubourg, S. Bailey, E. Balbinot, and et al., *SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way Galaxy, and Extra-Solar Planetary Systems*, *ArXiv e-prints* (Jan., 2011) [arXiv:1101.1529].

[25] A. Smette, J. Surdej, P. A. Shaver, C. B. Foltz, F. H. Chaffee, R. J. Weymann, R. E. Williams, and P. Magain, *A spectroscopic study of UM 673 A and B - On the size of Lyman-alpha clouds*, *ApJ* **389** (Apr., 1992) 39–62.

[26] J. Bechtold, A. P. S. Crotts, R. C. Duncan, and Y. Fang, *Spectroscopy of the double quasars Q1343+266A, B: A new determination of the size of Lyman-alpha forest absorbers*, *ApJL* **437** (Dec., 1994) L83–L86.

[27] N. Dinshaw, C. D. Impey, C. B. Foltz, R. J. Weymann, and F. H. Chaffee, *Common Lyman-alpha absorption toward the quasar pair Q1343+2640A, B: Evidence for large and quiescent clouds*, *ApJL* **437** (Dec., 1994) L87–L90.

[28] N. Dinshaw, C. B. Foltz, C. D. Impey, R. J. Weymann, and S. L. Morris, *Large size of Lyman-α gas clouds at intermediate redshifts*, *Nature* **373** (Jan., 1995) 223–225.

[29] M. Rauch and M. G. Haehnelt, *Omega_baryon and the geometry of intermediate-redshift Lyman alpha absorption systems*, *MNRAS* **275** (Aug., 1995) L76–L78, [astro-ph/].

[30] R. Cen, J. Miralda-Escude, J. P. Ostriker, and M. Rauch, *Gravitational collapse of small-scale structure as the origin of the Lyman-alpha forest*, *ApJL* **437** (Dec., 1994) L9–L12.

[31] Y. Zhang, P. Anninos, and M. L. Norman, *A Multispecies Model for Hydrogen and Helium Absorbers in Lyman-Alpha Forest Clouds*, *ApJL* **453** (Nov., 1995) L57+.

[32] L. Hernquist, N. Katz, D. H. Weinberg, and J. Miralda-Escudé, *The Lyman-Alpha Forest in the Cold Dark Matter Model*, *ApJL* **457** (Feb., 1996) L51+.

[33] J. Miralda-Escude, R. Cen, J. P. Ostriker, and M. Rauch, *The LY alpha Forest from Gravitational Collapse in the Cold Dark Matter + Lambda Model*, *ApJ* **471** (Nov., 1996) 582–+.

[34] D. H. Weinberg, N. Katz, and L. Hernquist, *Simulating Cosmic Structure Formation*, in *Origins* (C. E. Woodward, J. M. Shull, & H. A. Thronson Jr., ed.), vol. 148 of *Astronomical Society of the Pacific Conference Series*, pp. 21–+, 1998. astro-ph/.

[35] R. A. C. Croft, D. H. Weinberg, N. Katz, and L. Hernquist, *Recovery of the Power Spectrum of Mass Fluctuations from Observations of the Ly alpha Forest*, *ApJ* **495** (Mar., 1998) 44–+.

[36] L. Hui and N. Y. Gnedin, *Equation of state of the photoionized intergalactic medium*, *MNRAS* **292** (Nov., 1997) 27–+.

[37] M. S. Peeples, D. H. Weinberg, R. Davé, M. A. Fardal, and N. Katz, *Pressure support versus thermal broadening in the Lyman α forest - I. Effects of the equation of state on longitudinal structure*, *MNRAS* **404** (May, 2010) 1281–1294, [arXiv:0910.0256].

[38] N. Kaiser, *Clustering in real space and in redshift space*, *MNRAS* **227** (July, 1987) 1–21.

[39] P. McDonald, *Toward a Measurement of the Cosmological Geometry at $z \sim 2$: Predicting Lyα Forest Correlation in Three Dimensions and the Potential of Future Data Sets*, *ApJ* **585** (Mar., 2003) 34–51.

[40] A. Slosar, S. Ho, M. White, and T. Louis, *The Acoustic Peak in the Lyman Alpha Forest*, *JCAP* **0910** (2009) 019, [arXiv:0906.2414].

[41] M. White, A. Pope, J. Carlson, K. Heitmann, S. Habib, P. Fasel, D. Daniel, and Z. Lukic, *Particle Mesh Simulations of the Lyα Forest and the Signature of Baryon Acoustic Oscillations in the Intergalactic Medium*, *ApJ* **713** (Apr., 2010) 383–393, [arXiv:0911.5341].

[42] N. Y. Gnedin and L. Hui, *Probing the Universe with the Lyalpha forest - I. Hydrodynamics of the low-density intergalactic medium*, *MNRAS* **296** (May, 1998) 44–55.

[43] P. McDonald, J. Miralda-Escudé, M. Rauch, W. L. W. Sargent, T. A. Barlow, and R. Cen, *A Measurement of the Temperature-Density Relation in the Intergalactic Medium Using a New Lyα Absorption-Line Fitting Method*, *ApJ* **562** (Nov., 2001) 52–75.

[44] M. McQuinn, L. Hernquist, A. Lidz, and M. Zaldarriaga, *The Signatures of Large-scale Temperature Fluctuations in the Lyman-alpha Forest*, *ArXiv e-prints* (Oct., 2010) [arXiv:1010.5250].

[45] G. M. Williger, A. Smette, C. Hazard, J. A. Baldwin, and R. G. McMahon, *Evidence for Large-Scale Structure in the Lyα Forest at $z > 2.6$*, *ApJ* **532** (Mar., 2000) 77–87, [astro-ph/].

[46] E. Rollinde *et. al.*, *The correlation of the Lyman-alpha forest in close pairs and groups of high-redshift quasars: clustering of matter on scales 1-5 Mpc*, *Mon. Not. Roy. Astron. Soc.* **341** (2003) 1279, [astro-ph/0302145].

[47] G. D. Becker, W. L. W. Sargent, and M. Rauch, *Large-Scale Correlations in the Lyα Forest at $z = 3$-4*, *ApJ* **613** (Sept., 2004) 61–76, [astro-ph/].

[48] V. D'Odorico, M. Viel, F. Saitta, S. Cristiani, S. Bianchi, B. Boyle, S. Lopez, J. Maza, and P. Outram, *Tomography of the intergalactic medium with Lyα forests in close QSO pairs*, *MNRAS* **372** (Nov., 2006) 1333–1344, [astro-ph/].

[49] F. Coppolani, P. Petitjean, F. Stoehr, E. Rollinde, C. Pichon, S. Colombi, M. G. Haehnelt, B. Carswell, and R. Teyssier, *Transverse and longitudinal correlation functions in the intergalactic medium from 32 close pairs of high-redshift quasars*, *MNRAS* **370** (Aug., 2006) 1804–1816, [astro-ph/0605618].

[50] C. M. Casey, C. D. Impey, C. E. Petry, A. R. Marble, and R. Davé, *Pc 1643+4631A, b: the LYMAN-α Forest at the Edge of Coherence*, *AJ* **136** (July, 2008) 181–196, [arXiv:0804.2257].

[51] A. R. Marble, K. A. Eriksen, C. D. Impey, L. Bai, and L. Miller, *The Flux Auto- and Cross-Correlation of the Lyα Forest. I. Spectroscopy of QSO Pairs with Arcminute Separations and Similar Redshifts*, *ApJS* **175** (Mar., 2008) 29–47, [arXiv:0803.1851].

[52] D. Tytler, M. Gleed, C. Melis, A. Chapman, D. Kirkman, D. Lubin, P. Paschos, T. Jena, and A. P. S. Crotts, *Metal absorption systems in spectra of pairs of QSOs: how absorbers cluster around QSOs and other absorbers*, *MNRAS* **392** (Feb., 2009) 1539–1572.

[53] M. Cappetta, V. D'Odorico, S. Cristiani, F. Saitta, and M. Viel, *High-resolution spectroscopy of the 3D cosmic web with close QSO groups*, *MNRAS* **407** (Sept., 2010) 1290–1300, [arXiv:1004.0221].

[54] M. White, *The Ly-a forest*, in *The Davis Meeting On Cosmic Inflation*, Mar., 2003.

[55] P. McDonald and D. J. Eisenstein, *Dark energy and curvature from a future baryonic acoustic oscillation survey using the Lyman-α forest*, *Phys. Rev. D* **76** (Sept., 2007) 063009–+, [astro-ph/0607122].

[56] M. McQuinn and M. White, *On Estimating Lyman-alpha Forest Correlations between Multiple Sightlines*, *ArXiv e-prints* (Feb., 2011) [arXiv:1102.1752].

[57] M. White, M. Blanton, A. Bolton, D. Schlegel, J. Tinker, A. Berlind, L. da Costa, E. Kazin, Y. Lin, M. Maia, C. K. McBride, N. Padmanabhan, J. Parejko, W. Percival, F. Prada, B. Ramos, E. Sheldon, F. de Simoni, R. Skibba, D. Thomas, D. Wake, I. Zehavi, Z. Zheng, R. Nichol, D. P. Schneider, M. A. Strauss, B. A. Weaver, and D. H. Weinberg, *The Clustering*

of Massive Galaxies at z ~ 0.5 from the First Semester of BOSS Data, *ApJ* **728** (Feb., 2011) 126–+, [arXiv:1010.4915].

[58] D. P. Schneider, G. T. Richards, P. B. Hall, M. A. Strauss, S. F. Anderson, T. A. Boroson, N. P. Ross, Y. Shen, W. N. Brandt, X. Fan, N. Inada, S. Jester, G. R. Knapp, C. M. Krawczyk, A. R. Thakar, D. E. Vanden Berk, W. Voges, B. Yanny, D. G. York, N. A. Bahcall, D. Bizyaev, M. R. Blanton, H. Brewington, J. Brinkmann, D. Eisenstein, J. A. Frieman, M. Fukugita, J. Gray, J. E. Gunn, P. Hibon, Ž. Ivezić, S. M. Kent, R. G. Kron, M. G. Lee, R. H. Lupton, E. Malanushenko, V. Malanushenko, D. Oravetz, K. Pan, J. R. Pier, T. N. Price, D. H. Saxe, D. J. Schlegel, A. Simmons, S. A. Snedden, M. U. SubbaRao, A. S. Szalay, and D. H. Weinberg, *The Sloan Digital Sky Survey Quasar Catalog. V. Seventh Data Release*, *AJ* **139** (June, 2010) 2360–2373, [arXiv:1004.1167].

[59] N. P. Ross, A. D. Myers, E. S. Sheldon, C. Yeche, M. A. Strauss, J. A. K. Jo Bovy, G. T. Richards, E. Aubourg, M. R. Blanton, W. N. Brandt, W. C. Carithers, R. A. Croft, R. da Silva, K. Dawson, D. J. Eisenstein, J. F. Hennawi, S. Ho, D. W. Hogg, K.-G. Lee, B. Lundgren, R. G. McMahon, J. Miralda-Escudé, N. Palanque-Delabrouille, I. Pâris, P. Petitjean, M. M. Pieri, J. Rich, N. A. Roe, D. Schiminovich, D. J. Schlegel, D. P. Schneider, A. Slosar, N. Suzuki, J. L. Tinker, D. H. Weinberg, A. Weyant, M. White, and W. M. Wood-Vasey, *The SDSS-III Baryon Oscillation Spectroscopic Survey: Qusar Target Selection fot Data Release Nine*, *in preparation* (2011).

[60] G. T. Richards, A. D. Myers, A. G. Gray, R. N. Riegel, R. C. Nichol, R. J. Brunner, A. S. Szalay, D. P. Schneider, and S. F. Anderson, *Efficient Photometric Selection of Quasars from the Sloan Digital Sky Survey. II. ~1,000,000 Quasars from Data Release 6*, *ApJS* **180** (Jan., 2009) 67–83, [arXiv:0809.3952].

[61] C. Yèche, P. Petitjean, J. Rich, E. Aubourg, N. Busca, J. Hamilton, J. Le Goff, I. Paris, S. Peirani, C. Pichon, E. Rollinde, and M. Vargas-Magaña, *Artificial neural networks for quasar selection and photometric redshift determination*, *A&A* **523** (Nov., 2010) A14+.

[62] J. Kirkpatrick *et. al.*, *in preparation, 2011.*

[63] G. T. Richards, X. Fan, H. J. Newberg, M. A. Strauss, D. E. Vanden Berk, D. P. Schneider, B. Yanny, A. Boucher, S. Burles, J. A. Frieman, J. E. Gunn, P. B. Hall, Ž. Ivezić, S. Kent, J. Loveday, R. H. Lupton, C. M. Rockosi, D. J. Schlegel, C. Stoughton, M. SubbaRao, and D. G. York, *Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Quasar Sample*, *AJ* **123** (June, 2002) 2945–2975, [astro-ph/0202251].

[64] P. Noterdaeme, P. Petitjean, C. Ledoux, and R. Srianand, *Evolution of the cosmological mass density of neutral gas from Sloan Digital Sky Survey II - Data Release 7*, *A&A* **505** (Oct., 2009) 1087–1098, [arXiv:0908.1574].

[65] J. R. Trump, P. B. Hall, T. A. Reichard, G. T. Richards, D. P. Schneider, D. E. Vanden Berk, G. R. Knapp, S. F. Anderson, X. Fan, J. Brinkman, S. J. Kleinman, and A. Nitta, *A Catalog of Broad Absorption Line Quasars from the Sloan Digital Sky Survey Third Data Release*, *ApJS* **165** (July, 2006) 1–18, [astro-ph/].

[66] N. Suzuki, D. Tytler, D. Kirkman, J. M. O'Meara, and D. Lubin, *Predicting QSO Continua in the Lyα Forest*, *ApJ* **618** (Jan., 2005) 592–600.

[67] A. Font-Ribera *et. al.*, *in preparation, 2011.*

[68] Z. Zheng and J. Miralda-Escudé, *Self-shielding Effects on the Column Density Distribution of Damped Lyα Systems*, *ApJL* **568** (Apr., 2002) L71–L74.

[69] P. McDonald, U. Seljak, R. Cen, P. Bode, and J. P. Ostriker, *Physical effects on the Lyα forest flux power spectrum: damping wings, ionizing radiation fluctuations and galactic winds*, *MNRAS* **360** (July, 2005) 1471–1482.

[70] J. X. Prochaska and S. Herbert-Fort, *The Sloan Digital Sky Survey Damped Lyα Survey: Data*

Release 1, *PASP* **116** (July, 2004) 622–633.

[71] C. Péroux, M. Dessauges-Zavadsky, S. D'Odorico, T. Kim, and R. G. McMahon, *A homogeneous sample of sub-damped Lyman α systems - II. Statistical, kinematic and chemical properties*, *MNRAS* **345** (Oct., 2003) 480–496.

[72] C. Péroux, R. G. McMahon, L. J. Storrie-Lombardi, and M. J. Irwin, *The evolution of Ω_HI and the epoch of formation of damped Lyman α absorbers*, *MNRAS* **346** (Dec., 2003) 1103–1115.

[73] M. M. Pieri, S. Frank, D. H. Weinberg, S. Mathur, and D. G. York, *The Composite Spectrum of Strong Lyα Forest Absorbers*, *ApJL* **724** (Nov., 2010) L69–L73, [arXiv:1001.5282].

[74] A. Lewis and S. Bridle, *Cosmological parameters from CMB and other data: A Monte Carlo approach*, *Phys. Rev. D* **66** (Nov., 2002) 103511–+, [astro-ph/0205436].

[75] N. Ben Bekhti, P. Richter, T. Westmeier, and M. T. Murphy, *Ca II and Na I absorption signatures from extraplanar gas in the halo of the Milky Way*, *A&A* **487** (Aug., 2008) 583–594, [arXiv:0806.3204].

[76] P. McDonald and U. Seljak, *How to evade the sample variance limit on measurements of redshift-space distortions*, *JCAP* **10** (Oct., 2009) 7–+, [arXiv:0810.0323].

[77] K. Lee *et. al.*, *in preparation, 2011.*

[78] P. McDonald, *Clustering of dark matter tracers: Renormalizing the bias parameters*, *Phys. Rev. D* **74** (Nov., 2006) 103512–+, [astro-ph/0609413].

[79] P. McDonald and A. Roy, *Clustering of dark matter tracers: generalizing bias for the coming era of precision LSS*, *JCAP* **8** (Aug., 2009) 20–+, [arXiv:0902.0991].

[80] E. Komatsu *et. al.*, *Seven-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Interpretation*, arXiv:1001.4538.

[81] K. Lee and D. N. Spergel, *Threshold Clustering Functions and Thermal Inhomogeneities in the Lyman-Alpha Forest*, *ArXiv e-prints* (July, 2010) [arXiv:1007.3734].

[82] C. Alcock and B. Paczynski, *An evolution free test for non-zero cosmological constant*, *Nature* **281** (Oct., 1979) 358–+.

[83] P. McDonald and J. Miralda-Escudé, *Measuring the Cosmological Geometry from the Lyα Forest along Parallel Lines of Sight*, *ApJ* **518** (June, 1999) 24–31.

[84] L. Hui, A. Stebbins, and S. Burles, *A Geometrical Test of the Cosmological Energy Contents Using the LYalpha Forest*, *ApJL* **511** (Jan., 1999) L5–L8, [astro-ph/].

[85] H. Bi and A. F. Davidsen, *Evolution of Structure in the Intergalactic Medium and the Nature of the Ly alpha Forest*, *ApJ* **479** (Apr., 1997) 523–+, [astro-ph/].

A Appendix: Removing the mean of the forest

Consider a toy model in which a quasar has a constant continuum and we measure flux in pixels $i = 1 \dots N$:

$$f_i = \bar{f}(1 + \delta_i + \epsilon_i), \tag{A.1}$$

where δ_i is the underlying fluctuation field and ϵ_i our measurement error (we can rescale it by $\bar{f}$ without loss of generality). By fitting a continuum to the set of points and estimating the flux contrast, we actually estimate δ_i as:

$$\begin{aligned} \delta_i' = \frac{f_i - N^{-1}\sum_k f_i}{N^{-1}\sum_k f_i} &= \frac{\delta_i + \epsilon_i - N^{-1}\sum_k(\delta_k + \epsilon_k)}{1 + N^{-1}\sum_k(\delta_k + \epsilon_k)} \\ &\approx \left(\delta_i + \epsilon_i - N^{-1}\sum_k(\delta_k + \epsilon_k)\right)\left(1 - N^{-1}\sum_k(\delta_k + \epsilon_k)\right). \end{aligned} \tag{A.2}$$

In the approximation, we have used the fact that the mean fluctuation across the forest is much less than unity. Taking expectation value over noise, one gets

$$\langle \delta_i' \rangle = \delta_i - N^{-1}\sum_k \delta_k - \delta_i N^{-1}\sum_k \delta_k + N^{-2}\sum_{kl}\delta_k\delta_l - N^{-1}\sigma_i^2 + N^{-2}\sum_k \sigma_k^2 \tag{A.3}$$

where we assumed diagonal noise vector $\langle \epsilon_k \epsilon_l \rangle = \delta^K_{kl}\sigma^2_k$ (neglecting the terms from the denominator above).

This means that, after fitting for mean continuum, the estimator is not unbiased anymore and can, in principle, lead to change in large-scale bias[2].

We proceed to look at cross-correlation between two adjacent quasars with respective flux contrast measurements δ'^A and δ'^B which we, for simplicity, assume to have forests of equal length. Then, the trivial correlation function estimator gives

$$\langle \delta_i'^A \delta_j'^B \rangle = \langle \delta_i^A \delta_j^B \rangle - \frac{1}{N}\sum_k(\langle \delta_i^A \delta_k^B \rangle + \langle \delta_k^A \delta_j^B \rangle) + \frac{1}{N^2}\sum_k\sum_l \langle \delta_k^A \delta_l^B \rangle + \text{h. o. correlators} \tag{A.4}$$

(neglecting the terms from the denominator above).

Therefore for a given pair of pixels, the process of removing the mean component from the quasar results in measuring the true correlation function minus the appropriately averaged correlation function averaged over pixel pairs in the quasar.

In practice, one does not need to simulate the full geometry of the survey to calculate this effect; it is sufficient (as proven by tests on synthetic data) to assume that a typical correlation function is averaged over some distance Δr in the positions of both quasars:

$$\xi_F'(r_\perp, r_\parallel) = \xi_F(r_\perp, r_\parallel) - 2\frac{1}{\Delta r}\int_{-\Delta r/2}^{\Delta r/2} dr_1 \xi_F(r_\perp, r_\parallel + r_1) + \frac{1}{\Delta r^2}\iint_{-\Delta r/2}^{\Delta r/2} dr_1 dr_2 \xi_F(r_\perp, r_\parallel + r_1 - r_2) \tag{A.5}$$

[2] Note that this calculation is incomplete as we are missing the third-order terms which contribute at the same order when computing a correlation function.

where note that the $r_\parallel$ inside the integrals is not derived from Eq. A.4 – we are approximating the distribution of relative quasar redshifts by assuming that all quasars are at the same redshift in the $r_\parallel = 0$ case, and then assuming that the slightly different weightings of alignments for non-zero $r_\parallel$ lead effectively to a shift in alignment by exactly $r_\parallel$.

We use this simple formula to account for the overall affect of removing means from all spectra, with a single fitted Δr in each redshift bin, even though generally we could do a more careful spectrum-by-spectrum calculation using the pixel-pair weights, because mocks show that this approximation is good enough.

B Appendix: Errors of the trivial estimator

In this work we use the trivial correlation function estimator (we drop the subindex F in the flux correlation function in this Appendix to reduce clutter):

$$\bar{\xi}(r,\mu) = \frac{\sum_{\text{pairs } i,j} w_i w_j \delta_{f_i} \delta_{f_j}}{\sum_{\text{pairs } i,j} w_i w_j}, \tag{B.1}$$

It is clear that the expectation value of this estimator is the true correlation function, regardless of which weights w_i are used:

$$\left\langle \bar{\xi}(r,\mu) \right\rangle = \frac{\sum_{\text{pairs } i,j} w_i w_j \left\langle \delta_{fi} \delta_{fj} \right\rangle}{\sum_{\text{pairs } i,j} w_i w_j} = \left\langle \delta_{fi} \delta_{fj} \right\rangle \frac{\sum_{\text{pairs } i,j} w_i w_j}{\sum_{\text{pairs } i,j} w_i w_j} = \xi(r,\mu). \tag{B.2}$$

(where the correlation function is assumed to be constant within a bin). In other words, the estimator is un-biased. The co-variance of this estimator is given by (we use A and B to denote two (r,μ) pairs):

$$\left\langle \bar{\xi}(A) \bar{\xi}(B) \right\rangle = \frac{\sum_{\text{pairs } i,j \in A, \text{pairs } k,l\ \in B} w_i w_j w_k w_l \left\langle \delta_{fi} \delta_{fj} \delta_{fk} \delta_{fl} \right\rangle}{\sum_{\text{pairs } i,j \in A} w_i w_j \sum_{\text{pairs } i,j \in B} w_i w_j} \tag{B.3}$$

The 4-point term in brackets can be expanded using Wick's theorem, which yields three terms, the first of which is separable, giving

$$\left\langle \bar{\xi}(A) \bar{\xi}(B) \right\rangle = \xi(A)\xi(B) + \frac{\sum_{\text{pairs } i,j \in A, \text{pairs } k,l \in B} w_i w_j w_k w_l (\xi_{ik}\xi_{jl} + \xi_{il}\xi_{jk})}{\sum_{\text{pairs } i,j \in A} w_i w_j \sum_{\text{pairs } i,j \in B} w_i w_j} \tag{B.4}$$

and hence the covariance matrix of errors is given by

$$C_{AB} \left\langle \bar{\xi}(A) \bar{\xi}(B) \right\rangle - \xi(A)\xi(B) = \frac{\sum_{\text{pairs } i,j \in A, \text{pairs } k,l \in B} w_i w_j w_k w_l (\xi_{ik}\xi_{jl} + \xi_{il}\xi_{jk})}{\sum_{\text{pairs } i,j \in A} w_i w_j \sum_{\text{pairs } i,j \in B} w_i w_j}. \tag{B.5}$$

We use the raw measurement of the correlation function from the data for ξ here, including the noise contribution to the diagonal (in an averaged sense, not pixel-by-pixel). Strictly speaking this estimator is true only for Gaussian fields and the corrections to it are of the order of bispectrum.

C Appendix: Measuring forest metal absorption

A modified version of the approach described by [73] is used to measure metal absorption associated with the Lyman-α forest. We eliminate the requirement that stacked pixels be a local flux minimum, which was intended to limit the stacking of wings from stronger lines. This is preferable because our goal is not to measure metallicity, but to measure a signal in order to reproduce it. We combine the individual spectra using the arithmetic mean with a 3% outlier clipping and continuum fit to correct for uncorrelated absorption giving us a composite transmitted flux, F_c, of stacked systems.

We select pixels to stack by virtue of their normalized flux $F_n \equiv F/\bar{F}$, where the mean transmitted flux, $\bar{F}$, is determined using the method set out in § 4.2. Seven composite spectra were produced with $F_n < 0.4$ (above which the metal signal was negligible). We retain the requirement from [73] that pixels be 0.5σ from saturation (a standard choice in pixel optical depth techniques) in order to obtain a clean measure and minimize the contribution of LLS/DLAs. In each of our composite spectra, we measure F_c at line center for 7 metal transitions: Si II (1193Å), Si III (1207Å), N V (1239, 1242Å), Si II (1260Å), O I (1302Å), Si II (1304Å). Since we do not set a requirement that all stacked absorption be Lyman-α , we allow that some lines in the composite arise from stacking metal lines. The only resolved contamination of this sort arises from stacking Si III lines. We have measured F_c at line center for 7 wavelengths where this 'shadow' signal would be present (see [73] for details). We include this shadow signal in the construction of mocks as if it came from, e.g., an additional metal line at $\sim$ 1225Å, even though this implies that the original underlying mock field represents not just a hydrogen field, but also an identical Si III field unphysically offset by $\sim 20\,h^{-1}$ Mpc in the radial direction. We don't think this affects our results.

This process provides a look-up table of 7 Lyman-α line strengths and normalized flux decrements ($D_c \equiv 1 - F_c$) at 14 fixed spectral locations. This is shown in Table 3. It should be noted that we do not limit ourselves to high-significance lines (as in dedicated metal line studies). Instead we include all metal lines that may introduce contamination, and are seen in some composite spectra. We constrain the metal signal to be positive, which biases our results upward (e.g., if a line did not really exist, we would on average add one), but we do not think this affects our results, because generally the non-detections have fairly tight upper limits.

No significant evolution in the metal line strengths is seen, but as the size of the BOSS survey grows we will revisit the analysis. Here we use the full redshift range used in this flux correlation analysis.

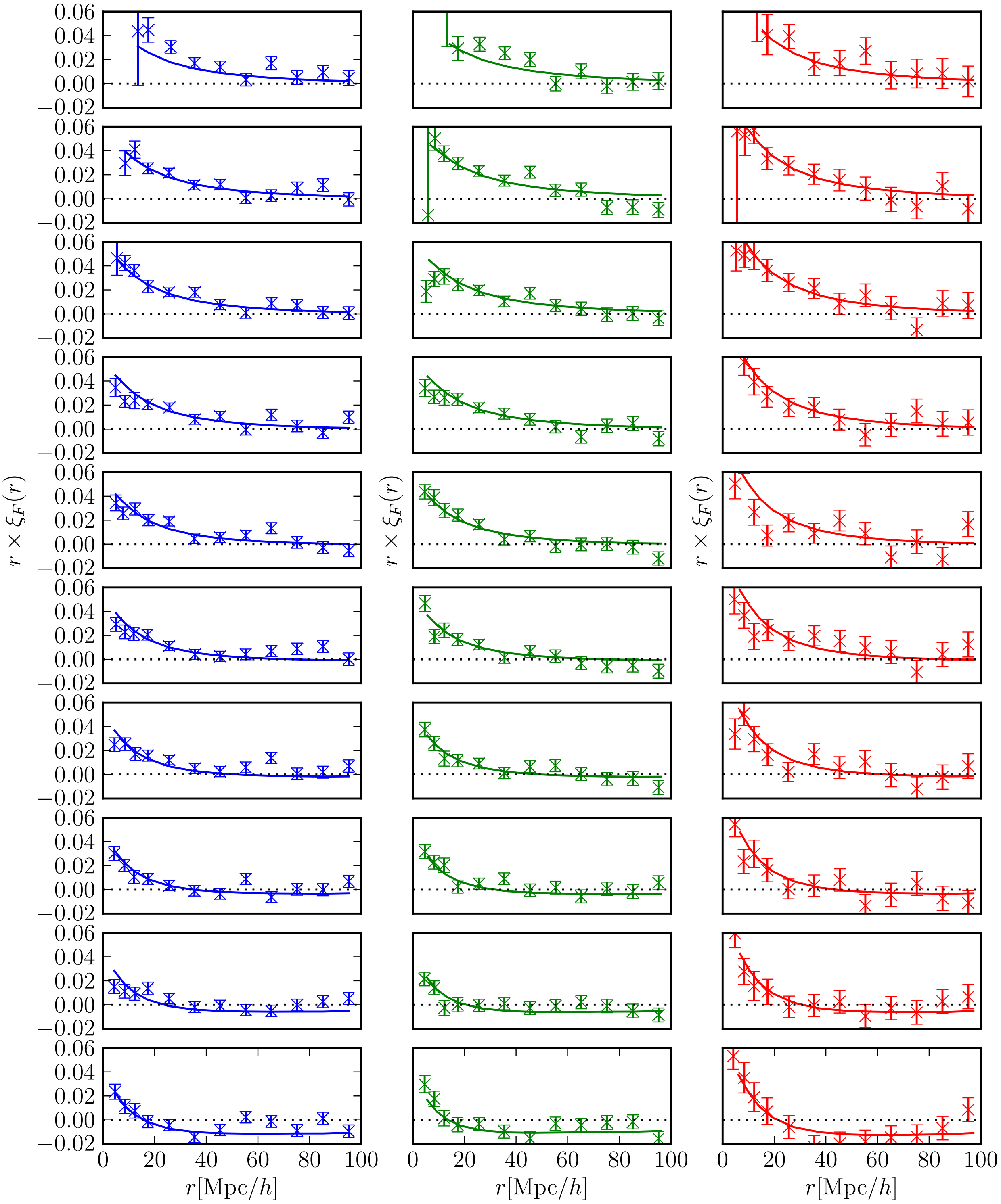

Figure 16. Measurements of the observed data. Columns correspond to the three redshift bins we use, with increasing redshift from left to right. The ten rows correspond to the ten bins of μ, increasing from top to bottom. In each plot we show the measured ξ_F as a function of separation for that particular redshift and μ bin. The best-fit linear theory is over-plotted to guide the eye. The measurements are correlated and hence one should not evaluate "χ^2 by eye".

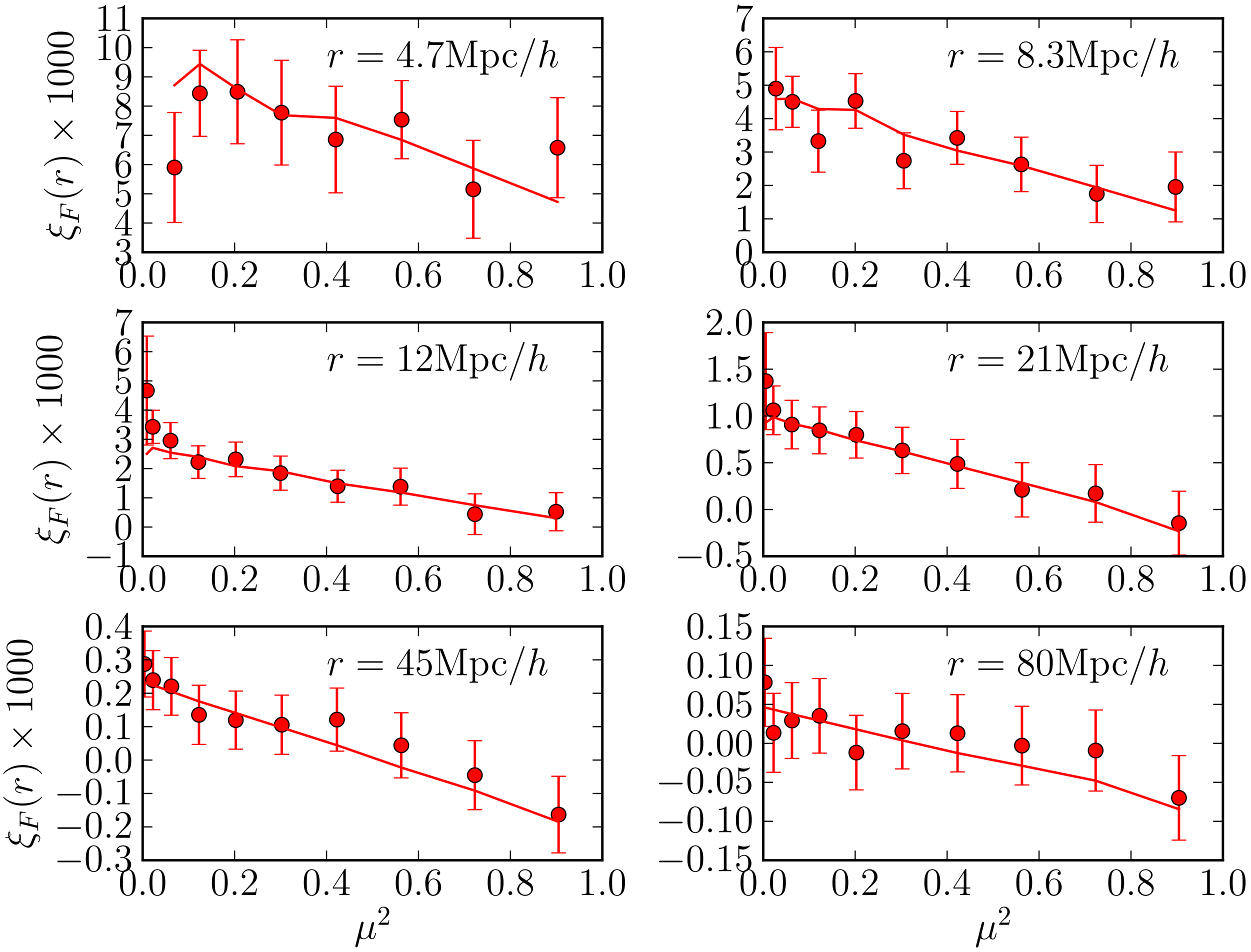

Figure 17. Measurements of the observed data. Each panel corresponds to redshift-averaged data at a certain radius as a function of μ. We also plot the best-fit linear model to guide the eye. Measurements are correlated and hence one should not evaluate "χ^2 by eye".

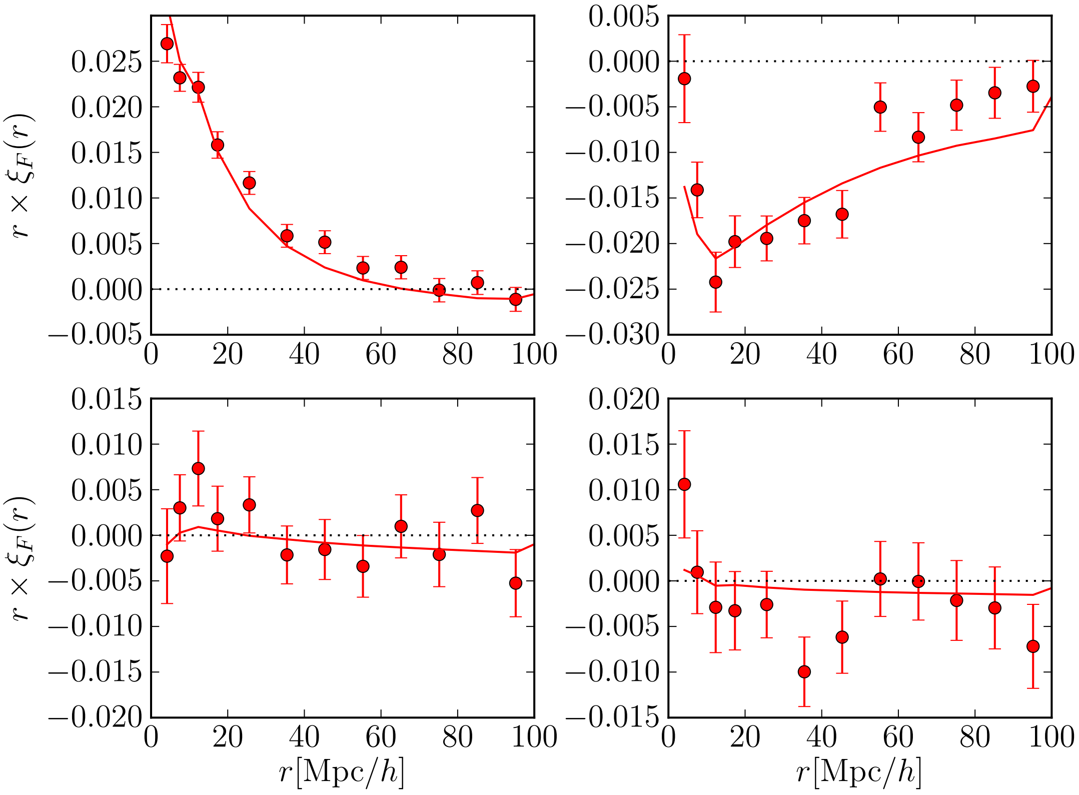

Figure 18. Results of Figure 16 converted to multipoles. The four panels correspond to the redshift-averaged monopole, quadrupole (top row), hexadecapole and $\ell = 6$ moment (bottom row). Lines are best-fit theory.

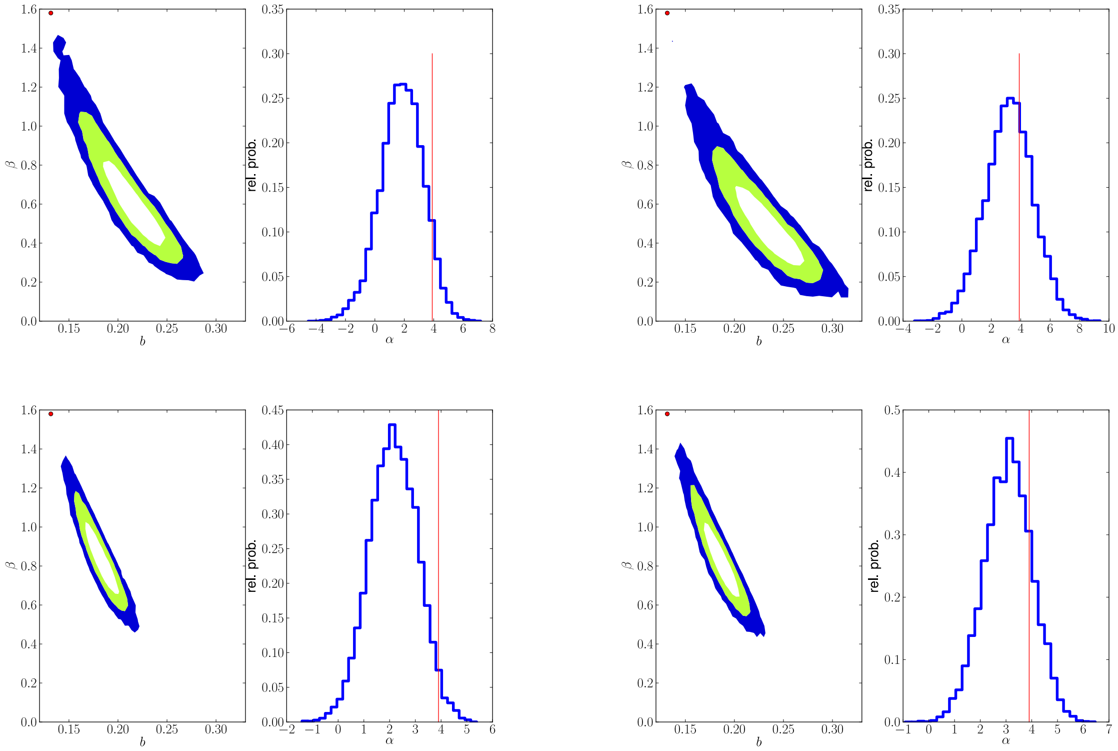

Figure 19. Fits to the real data. The upper panels correspond to the data fitted using points with separations $r > 20h^{-1}$Mpc while the lower panel is for fits using points with $r > 10h^{-1}$Mpc. The left-hand-side plots is for the default dataset, while the right hand side plot is for data that include quasars flagged as harboring DLAs by the FPG. The red-point corresponds to the value that was used in the creation of the synthetic datasets.

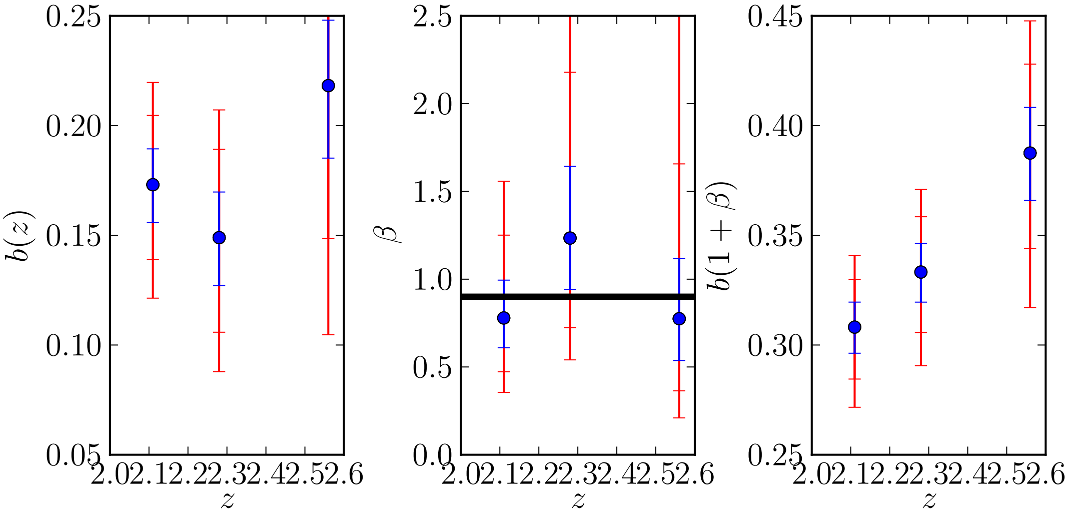

Figure 20. In this plot we show fits to the real data when the b and β parameters are allowed to be a function of redshift. We plot the 1,2 and 3 σ error bars. The bias in this plot is with respect to the fiducial cosmological model with $\sigma_8 = 0.8$ at the redshift of interest, therefore the numbers cannot be directly compared with the fitted α,b parameters. We plot the value of β determined from the overall fit as a black solid line. All fits in this figure are limited to $r > 10h^{-1}$Mpc.

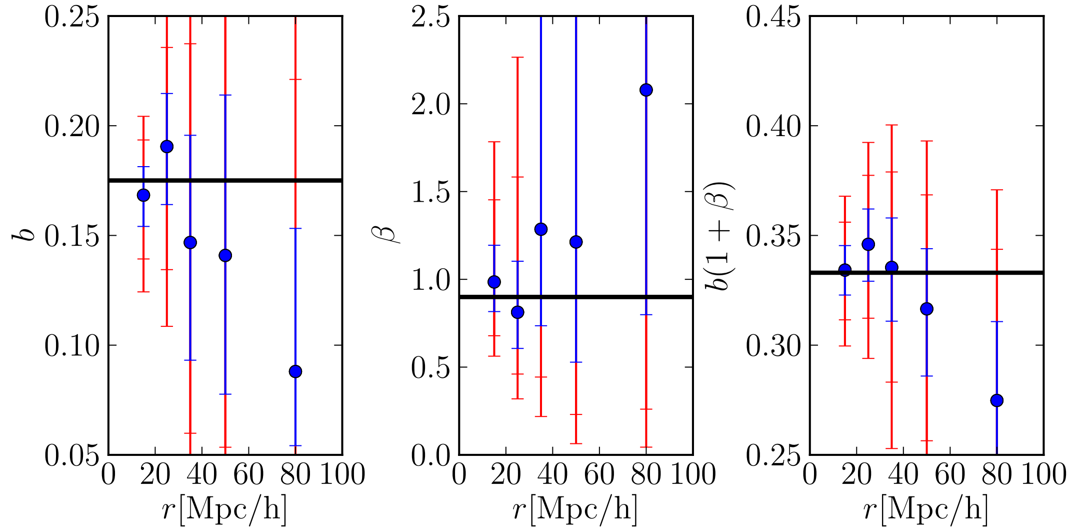

Figure 21. Results of fits of the real data to the b and β parameters when they are allowed to be a function of scale. We plot constraints on b, β and $b(1+\beta)$, and their 1,2 and 3 σ error bars measured from the corresponding percentiles of MCMC chains. Note that β and $b(1+\beta)$ have flat priors on them and that b is a derived parameters. The solid black thick lines correspond to the best fit parameters determined from fitting $r > 10h^{-1}$Mpc points.

			($10^3 D_c$) from correlation with Lyman-α							($10^3 D_c$) from correlation with Si III						
Species			Si II	Si III	N V	N V	Si II	O I	Si II	Si II	Lyman-α	N V	N V	Si II	O I	Si II
λ			1193Å	1207Å	1239Å	1242Å	1260Å	1302Å	1304Å	1202Å	1225Å	1248Å	1252Å	1270Å	1312Å	1314Å
Stacked F_n																
Min	Max	Mean														
0.02	0.1	0.074	40 ± 10	123 ± 8	8 ± 7	5 ± 8	50 ± 8	40 ± 10	30 ± 10	0 ± 6	13 ± 8	1 ± 9	0 ± 7	1 ± 9	10 ± 10	0 ± 10
0.1	0.15	0.128	8 ± 9	53 ± 7	11 ± 6	0 ± 6	10 ± 6	0 ± 8	12 ± 9	2 ± 6	17 ± 6	0 ± 6	0 ± 7	12 ± 7	21 ± 9	0 ± 8
0.15	0.2	0.177	5 ± 6	20 ± 7	5 ± 5	0 ± 4	6 ± 5	13 ± 6	12 ± 5	1 ± 6	14 ± 4	2 ± 5	0 ± 4	9 ± 6	7 ± 6	0 ± 6
0.2	0.25	0.227	0 ± 6	8 ± 4	3 ± 4	3 ± 4	3 ± 4	1 ± 4	0 ± 5	4 ± 5	10 ± 3	3 ± 4	0 ± 4	4 ± 3	4 ± 6	10 ± 5
0.25	0.3	0.276	0 ± 5	0 ± 4	5 ± 3	3 ± 3	0 ± 3	1 ± 4	4 ± 4	0 ± 5	10 ± 3	3 ± 3	0 ± 3	3 ± 4	0 ± 5	6 ± 5
0.3	0.35	0.326	0 ± 5	1 ± 4	2 ± 3	1 ± 3	0 ± 3	0 ± 3	0 ± 4	2 ± 4	5 ± 3	2 ± 3	0 ±3	0 ± 3	0 ± 4	0 ± 4
0.35	0.4	0.376	0 ± 4	0 ± 4	0 ± 3	1 ± 2	1 ± 3	0 ± 3	3 ± 3	0 ± 4	10 ± 2	0 ± 3	3 ± 3	2 ± 3	2 ± 4	5 ± 3

Table 3. The flux decrement at line center, D_c, measured in the composite spectra of absorbers within the Lyman-α forest as described in Appendix C. Signal measured as a result of stacking both Lyman-α and Si III is shown. The correlated signal due to 8 transitions are measured.

References

J. Bechtold, A. P. S. Crotts, R. C. Duncan, and Y. Fang. Spectroscopy of the double quasars Q1343+266A, B: A new determination of the size of Lyman-alpha forest absorbers. *ApJL*, 437:L83–L86, December 1994.

R. Cen, J. Miralda-Escude, J. P. Ostriker, and M. Rauch. Gravitational collapse of small-scale structure as the origin of the Lyman-alpha forest. *ApJL*, 437:L9–L12, December 1994.

R. A. C. Croft, D. H. Weinberg, N. Katz, and L. Hernquist. Recovery of the Power Spectrum of Mass Fluctuations from Observations of the Ly alpha Forest. *ApJ*, 495: 44–+, March 1998.

R. A. C. Croft, D. H. Weinberg, M. Pettini, L. Hernquist, and N. Katz. The Power Spectrum of Mass Fluctuations Measured from the LYalpha Forest at Redshift Z=2.5. *ApJ*, 520:1–23, July 1999.

N. Dinshaw, C. D. Impey, C. B. Foltz, R. J. Weymann, and F. H. Chaffee. Common Lyman-alpha absorption toward the quasar pair Q1343+2640A, B: Evidence for large and quiescent clouds. *ApJL*, 437:L87–L90, December 1994.

N. Dinshaw, C. B. Foltz, C. D. Impey, R. J. Weymann, and S. L. Morris. Large size of Lyman-α gas clouds at intermediate redshifts. *Nature*, 373:223–225, January 1995. doi: 10.1038/373223a0.

D. J. Eisenstein, D. H. Weinberg, E. Agol, H. Aihara, C. Allende Prieto, S. F. Anderson, J. A. Arns, E. Aubourg, S. Bailey, E. Balbinot, and et al. SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way Galaxy, and Extra-Solar Planetary Systems. *ArXiv e-prints*, January 2011.

M. Fukugita, T. Ichikawa, J. E. Gunn, M. Doi, K. Shimasaku, and D. P. Schneider. The Sloan Digital Sky Survey Photometric System. *AJ*, 111:1748–+, April 1996. doi: 10.1086/117915.

C. M. Hirata. Tidal alignments as a contaminant of redshift space distortions. *MNRAS*, 399:1074–1087, October 2009. doi: 10.1111/j.1365-2966.2009.15353.x.

L. Jiang, X. Fan, R. J. Cool, D. J. Eisenstein, I. Zehavi, G. T. Richards, R. Scranton, D. Johnston, M. A. Strauss, D. P. Schneider, and J. Brinkmann. A Spectroscopic Survey of Faint Quasars in the SDSS Deep Stripe. I. Preliminary Results from the Co-added Catalog. *AJ*, 131:2788–2800, June 2006. doi: 10.1086/503745.

N. Kaiser. Clustering in real space and in redshift space. *MNRAS*, 227:1–21, July 1987.

P. McDonald. Toward a Measurement of the Cosmological Geometry at $z \sim 2$: Predicting Lyα Forest Correlation in Three Dimensions and the Potential of Future Data Sets. *ApJ*, 585:34–51, March 2003.

P. McDonald and D. J. Eisenstein. Dark energy and curvature from a future baryonic acoustic oscillation survey using the Lyman-α forest. *Phys. Rev. D*, 76(6):063009–+, September 2007. doi: 10.1103/PhysRevD.76.063009.

P. McDonald, J. Miralda-Escudé, M. Rauch, W. L. W. Sargent, T. A. Barlow, R. Cen, and J. P. Ostriker. The Observed Probability Distribution Function, Power Spectrum, and Correlation Function of the Transmitted Flux in the Lyα Forest. *ApJ*, 543:1–23, November 2000.

P. McDonald, U. Seljak, R. Cen, P. Bode, and J. P. Ostriker. Physical effects on the Lyα forest flux power spectrum: damping wings, ionizing radiation fluctuations and galactic winds. *MNRAS*, 360:1471–1482, July 2005. doi: 10.1111/j.1365-2966.2005.09141.x.

P. McDonald, U. Seljak, S. Burles, D. J. Schlegel, D. H. Weinberg, R. Cen, D. Shih, J. Schaye, D. P. Schneider, N. A. Bahcall, J. W. Briggs, J. Brinkmann, R. J. Brunner, M. Fukugita, J. E. Gunn, Ž. Ivezić, S. Kent, R. H. Lupton, and D. E. Vanden Berk. The Lyα Forest Power Spectrum from the Sloan Digital Sky Survey. *ApJS*, 163:80–109, March 2006. doi: 10.1086/444361.

M. McQuinn and M. White. On Estimating Lyman-alpha Forest Correlations between Multiple Sightlines. *ArXiv e-prints*, February 2011.

P. J. E. Peebles. *The large-scale structure of the universe*. Research supported by the National Science Foundation. Princeton, N.J., Princeton University Press, 1980. 435 p., 1980.

A. Pontzen, F. Governato, M. Pettini, C. M. Booth, G. Stinson, J. Wadsley, A. Brooks, T. Quinn, and M. Haehnelt. Damped Lyman α systems in galaxy formation simulations. *MNRAS*, 390:1349–1371, November 2008. doi: 10.1111/j.1365-2966.2008.13782.x.

W. H. Press and P. Schechter. Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation. *ApJ*, 187:425–438, February 1974.

J. X. Prochaska, S. Herbert-Fort, and A. M. Wolfe. The SDSS Damped Lyα Survey: Data Release 3. *ApJ*, 635:123–142, December 2005. doi: 10.1086/497287.

M. Rauch. The Lyman Alpha Forest in the Spectra of QSOs. *ARAA*, 36:267–316, 1998. doi: 10.1146/annurev.astro.36.1.267.

Nicholas P. Ross, Adam D. Myers, Erin S. Sheldon, Christophe Yeche, Michael A. Strauss, Jessica A. Kirkpatrick Jo Bovy, Gordon T. Richards, Eric Aubourg, Michael R. Blanton, W. N. Brandt, William C. Carithers, Rupert A.C. Croft, Robert da Silva, Kyle Dawson, Daniel J. Eisenstein, Joseph F. Hennawi, Shirley Ho, David W. Hogg, Khee-Gan Lee, Britt Lundgren, Richard G. McMahon, Jordi Miralda-Escudé, Nathalie Palanque-Delabrouille, Isabelle Pâris, Patrick Petitjean, Matthew M. Pieri, James Rich, Natalie A. Roe, David Schiminovich, David J. Schlegel, Donald P. Schneider, Anže Slosar, Nao Suzuki, Jeremy L. Tinker, David H. Weinberg, Anya Weyant, Martin White, and W. Michael Wood-Vasey. The SDSS-III Baryon Oscillation Spectroscopic Survey: Qusar Target Selection fot Data Release Nine. *in preparation*, 2011.

R. K. Sheth and G. Tormen. Large-scale bias and the peak background split. *MNRAS*, 308:119–126, September 1999.

A. Slosar, A. Font-Ribera, M. M. Pieri, J. Rich, J.-M. Le Goff, É. Aubourg, J. Brinkmann, N. Busca, B. Carithers, R. Charlassier, M. Cortês, R. Croft, K. S. Dawson, D. Eisenstein, J.-C. Hamilton, S. Ho, K.-G. Lee, R. Lupton, P. McDonald, B. Medolin, J. Miralda-Escudé, A. D. Myers, R. C. Nichol, N. Palanque-Delabrouille, I. Pâris, P. Petitjean, Y. Piškur, E. Rollinde, N. P. Ross, D. J. Schlegel, D. P. Schneider, E. Sheldon, B. A. Weaver, D. H. Weinberg, C. Yeche, and D. G. York. The Lyman-alpha forest in three dimensions: measurements of large scale flux correlations from BOSS 1st-year data, April 2011.

Anze Slosar, Shirley Ho, Martin White, and Thibaut Louis. The Acoustic Peak in the Lyman Alpha Forest. *JCAP*, 0910:019, 2009. doi: 10.1088/1475-7516/2009/10/019.

A. Smette, J. Surdej, P. A. Shaver, C. B. Foltz, F. H. Chaffee, R. J. Weymann, R. E. Williams, and P. Magain. A spectroscopic study of UM 673 A and B - On the size of Lyman-alpha clouds. *ApJ*, 389:39–62, April 1992. doi: 10.1086/171187.

N. Suzuki, D. Tytler, D. Kirkman, J. M. O'Meara, and D. Lubin. Predicting QSO Continua in the Lyα Forest. *ApJ*, 618:592–600, January 2005. doi: 10.1086/426062.

M. Tegmark, A. N. Taylor, and A. F. Heavens. Karhunen-Loeve Eigenvalue Problems in Cosmology: How Should We Tackle Large Data Sets? *ApJ*, 480:22+, May 1997.

D. H. Weinberg and S. Cole. Non-Gaussian fluctuations and the statistics of galaxy clustering. *MNRAS*, 259:652–694, December 1992.

M. White, M. Blanton, A. Bolton, D. Schlegel, J. Tinker, A. Berlind, L. da Costa, E. Kazin, Y.-T. Lin, M. Maia, C. K. McBride, N. Padmanabhan, J. Parejko, W. Percival, F. Prada, B. Ramos, E. Sheldon, F. de Simoni, R. Skibba, D. Thomas, D. Wake, I. Zehavi, Z. Zheng, R. Nichol, D. P. Schneider, M. A. Strauss, B. A. Weaver, and D. H. Weinberg. The Clustering of Massive Galaxies at z ˜ 0.5 from the First Semester of BOSS Data. *ApJ*, 728:126–+, February 2011. doi: 10.1088/0004-637X/728/2/126.

J. S. B. Wyithe. A method to measure the mass of damped Lyα absorber host galaxies using fluctuations in 21-cm emission. *MNRAS*, 388:1889–1898, August 2008. doi: 10.1111/j.1365-2966.2008.13546.x.

J. Yoo, D. H. Weinberg, J. L. Tinker, Z. Zheng, and M. S. Warren. Extending Recovery of the Primordial Matter Power Spectrum. *ApJ*, 698:967–985, June 2009. doi: 10.1088/0004-637X/698/2/967.

D. G. York et al. The Sloan Digital Sky Survey: Technical Summary. *AJ*, 120:1579–1587, September 2000.

Z. Zheng and J. Miralda-Escudé. Self-shielding Effects on the Column Density Distribution of Damped Lyα Systems. *ApJL*, 568:L71–L74, April 2002.