Dissertation

submitted to the

Combined Faculties of the Natural Sciences and Mathematics

of the Ruperto-Carola-University of Heidelberg. Germany

for the degree of

Doctor of Natural Sciences

Put forward by

Alberto Rorai

born in: San Daniele del Friuli (Italy)

Oral examination: 14/11/2014

Measuring the Small Scale Structure of the Intergalactic Medium

Referees: Joseph F. Hennawi

Volker Springel

Topic in German: Die kleinskalige Struktur des intergalaktischen Mediums ist grundlegend für das Verständnis von Kosmologie und Strukturbildung. Obwohl die Baryonen den Fluktuationen der dunklen Matterie auf Skalen in der Größenordnung von Megaparsec folgen, werden auf kleinen Skalen (~100 kpc) die Gaspertubationen durch hyrdodynamische Gleichungen reguliert. Es wird angenommen, dass sie unterhalb einer charaktersistischen Längenskala aufgrund von Druckgradienten unterdrückt werden, analog zur klassischen Jeanslänge. Der Wert der Jeansfilterlänge λ_J wird festgelegt durch ein Gleichgweicht zwischen Druck und Gravitationskräften und hat grundlegende kosmologische Anwendungen. Erstens liefert es einen thermischen Indiz für die zugeführte Wärme von unltravioletten Photonen während der Reionisation und bestimmt somit die thermische Geschichte des Universums. Zweitens bestimmt es die Verklumpung des IGM und die minimale gravitative Masse für den Kollaps des IGM, die eine zentrale Rolle in der Galaxienentstehung und Reionisation spielt. Prinzpiell kann Jeansglättung durch rotverschobene Lyman-α Absorptionslinien in Spektren von hoch rotverschobenen Quasaren nachgewiesen werden. Leider ist dies extrem schwierig, da die Auswirkungen des thermischen Dopplereffektes von Lyman-α Linien entlang der Beobachtungsrichtung von der Druckverbreiterung nicht klar zu trennen sind.
In dieser Arbeit zeige ich explizit, welche Entartungen zwischen den thermischen Parametern auftreten, wenn ausschließlich Beobachtungen entlang einer Sichtlinie möglich sind. Dafür habe ich einen stabilen statistischen Alogrithmus basierend auf Gaussprozessen und Markov Chain Monte Carlo Methoden entworfen, der auf einem Gitter eines semianalytischen Modells des IGM beruht. Ich führe dann eine neue Methode zum Messen der Jeanslänge ein, indem ich die transverse Kohärenz in Spektren benachbarter Quasarenpaare berechne (transverser Abstand < 1 Mpc). Diese Methode basiert auf der Phasendifferenz homologer Fouriermoden in dem Lyman-α Wald von Quasarenpaaren. Ich beweise, dass dies maximal empfindlich zu λ_J ist und nur schwach von anderen Parametern abhängt. Die verfügbare Stichprobe von Quasarenpaaren wird unter sorgfältiger Kalibration des Rauschens, der Auflösung und anderer möglicher systematischer Effekte ausgewertet. Unsere neue Methode auf diesen Datensatz angewendet gibt die erste Messung der Filterlänge des IGM. Ein erster Vergleich unserer Ergebnisse mit hydrodynamischen Simultaionen lässt darauf schließen, dass die vom thermischen Standardmodell des IGM vorrausgesagte Filterlänge signifikant höher ist als beobachtet. Dies motiviert weitere theoretische Studien zum Verständins dieser Diskrepanz.

Topic in English: The small-scale structure of the intergalactic medium (IGM) is fundamental to our understanding of cosmology and structure formation. Although the baryons trace dark matter fluctuations on megaparsec scales, on small scales (~100 kpc), gas perturbations are regulated by hydrodynamics and they are thought to be suppressed by pressure below a characteristic *filtering scale* λ_J, analogous to the classic Jeans scale. The value of this Jeans filtering scale is set by the interplay between pressure support and gravity across the cosmic history, and has fundamental cosmological implications. First it provides a thermal record of heat injected by ultraviolet photons during cosmic reionization events, and thus constraints the thermal and reionization history of the universe. Second, it determines the clumpiness of the IGM and the minimum mass for gravitational collapse from the IGM, playing a pivotal role in galaxy formation and reionization. In principle, the sign of Jeans smoothing could be probed by the redshifted Lyman-α absorption lines in the spectra of high-redshift quasars (the Lyman-α forest). Unfortunately, this is extremely challenging to do because the thermal Doppler broadening of Lyman- α lines along the observing direction is highly degenerate with pressure smoothing.
In this work, I explicitly show what degeneracies hold among the thermal parameters of the IGM when only line-of-sight observations are possible. For this purpose, I devised a rigorous statistical algorithm based on Gaussian processes and Markov-Chain Monte Carlo methods, trained on a grid of semianalytical models of the IGM. I then introduce a novel method able to measure the Jeans scale by estimating the transverse coherence in the spectra of close quasar pairs (transverse separation < 1 Mpc). This method is based on the phase differences of homologous Fourier modes in the Lyman-α forests of quasar pairs, and I prove that it is maximally sensitive to λ_J and only weakly dependent on the other considered parameters. The available sample of quasar pairs is analyzed, after careful calibration of noise, resolution, and other possible systematics. Our new method applied to this dataset provides the first measurement of the filtering scale of the intergalactic medium. A first comparison of our findings with hydrodynamical simulations suggests that the filtering scale predicted by the standard thermal models of the IGM is significantly higher than what we observe, motivating further theoretical studies to understand this discrepancy.

Universität Heidelberg

Doctoral Thesis

Measuring the Small Scale Structure of the Intergalactic Medium

Author:
Alberto Rorai

Supervisor:
Dr. Joseph Hennawi

A thesis submitted in fulfilment of the requirements
for the degree of Doctor in Astronomy

September 2014

Declaration of Authorship

I, Alberto Rorai, declare that this thesis titled, 'Measuring the Small Scale Structure of the Intergalactic Medium' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

Overview

In this manuscript I present the bulk of the work that I have conducted during my PhD under the supervision of Joseph F. Hennawi at the Max-Planck-Institut für Astronomie.

The initial goal of the project was to understand whether a recently discovered sample of quasar pairs could be used to probe the small-scale structure of the intergalactic medium (IGM), by studying the transverse coherence of the redshifted Lyα absorption in quasar spectra. The scientific motivations behind this objective are numerous. It opens the possibility of studying the structure evolution in the quasi-linear regime at the smallest length ever reached, which could be sensitive to unconstrained aspects of the cosmological models. In this work we focus on the relation with reionization and with the thermal evolution of the IGM: the pressure of the heated and ionized gas is expected to quench the growth of density perturbation below a characteristic scale called *Jeans scale*, or *filtering scale* (we will use the term as synonyms throughout the manuscript). This scale, although theoretically predicted, has never been constrained, and it may provide precious insights on galaxy formation and on the early stage of the reionization (a broader discussion is provided in chapter 1). This work represent the first attempt of measuring it at the redshifts of the Lyα forest.

The project has been carried on in two stages.

In the first stage we explored theoretically the sensitivity of the Lyα forest to the parameters that describe the thermal state of the IGM and in particular on the Jeans scale. We developed an algorithm that enables a systematic study of the sensitivity and degeneracies of Lyα-forest statistics with respect to the thermal parameter of the IGM, based on a set of semianalytical models. I describe this method and the models on which it relies in chapter 2. We then devised a new statistic specifically tailored to extract transverse-coherence information from quasar pairs. This statistic is based on the phase differences of homologous Fourier modes of the Lyα forests of two companion quasars, and we show in chapter 3 that it is maximally sensitive to the Jeans scale and practically insensitive on the other parameters that we analyze.

In the second part we applied the phase-difference method to the observed sample of quasar pairs, in the attempt of constraining the filtering scale of the IGM at redshift $2 < z < 3$. The main challenge in doing that was a proper treatment of noise, resolution and of the systematics (chapter 3), as well as understanding the exact meaning of the measured filtering scale and the extent to which our DM-based model could be trusted (chapter 6). The results we achieved, presented in chapter 5, indicate that the Jeans filtering scale is significantly smaller than what hydrodynamic simulation predicts for standard assumptions on the thermal history. The potentially controversial consequences of this finding demand further consideration of the possible systematic that could bias our measurement, and motivates a deeper theoretical exploration of the nature of the filtering scale and in particular of its relation with the thermal history.

The first three chapters present material that I have published in [Rorai et al., 2013], slightly adapted and reorganized to be inserted in this thesis, while the rest is unpublished. The work described in chapters 1-5 represent my personal contribution, except when I explicitly report results or methods from other studies. Chapter 6 contains results achieved in our research group in the past month, in particular in collaboration with Jose Oñorbe and Girish Kulkarni, in which I have been actively involved. I personally conducted the test described in § 6.2.2 to validate the calibration of my measurement with dark-matter simulations. I contributed to the definition of filtering scale of the IGM based on the Lyα absorption in 3d (§ 6.2), which will be published in Kulkarni et al. (in prep.) and on the fitting procedure described in § 6.3 (to be published in Oñorbe et al., in prep.).

Contents

Chapter 1

Introduction

The imprint of redshifted Lyman-α (Lyα) forest absorption on the spectra of distant quasars provides an exquisitely sensitive probe of the distribution of baryons in the intergalactic medium (IGM) at large cosmological lookback times. Among the remarkable achievements of modern cosmology is the ability of cosmological hydrodynamical simulations to explain the origin of this absorption pattern, and reproduce its statistical properties to percent level accuracy [e.g. Cen et al., 1994, Miralda-Escudé et al., 1996, Rauch, 1998]. But the wealth of information which can be gathered from the Lyα forest is far from being exhausted. The thermal state of the baryons in the IGM reflects the integrated energy balance of heating — due to the collapse of cosmic structures, radiation, and possibly other exotic heat sources — and cooling due to the expansion of the Universe [e.g. Hui & Gnedin, 1997, Hui & Haiman, 2003, Meiksin, 2009, Miralda-Escudé & Rees, 1994]. Cosmologists still do not understand how the interplay of these physical processes sets the thermal state of the IGM, nor has this thermal state been precisely measured.

There is ample observational evidence that ultraviolet radiation emitted by the first star-forming galaxies ended the 'cosmic dark ages' ionizing hydrogen and singly ionizing helium at $z \sim 10$ [e.g. Barkana & Loeb, 2001, Ciardi & Ferrara, 2005, Fan et al., 2006, Zaroubi, 2013]. A second and analogous reionization episode is believed to have occurred at later times $z \sim 3-4$ [Croft et al., 1997, Jakobsen et al., 1994, Madau & Meiksin, 1994, Reimers et al., 1997], when quasars were sufficiently abundant to supply the hard photons necessary to doubly ionized helium. The most recent observations from HST/COS provide tentative evidence for an extended He II reionization from $z \sim 2.7-4$ [Furlanetto & Dixon, 2010, Shull et al., 2010, Worseck et al., 2011, Worseck et al. 2013, in preparation], with a duration of $\sim 1\,\mathrm{Gyr}$, longer than naively expected. Cosmic reionization events are watersheds in the thermal history of the Universe, photoheating

the IGM to tens of thousands of degrees. Because cooling times in the rarefied IGM gas are long, memory of this heating is retained [Haehnelt & Steinmetz, 1998, Hui & Gnedin, 1997, Hui & Haiman, 2003, Miralda-Escudé & Rees, 1994, Theuns et al., 2002a,b]. Thus an empirical characterization of the IGMs thermal history constrains the nature and timing of reionization.

From a theoretical perspective, the impact of reionization events on the thermal state of the IGM is poorly understood. Radiative transfer simulations of both hydrogen [Bolton et al., 2004, Iliev et al., 2006, Tittley & Meiksin, 2007a] and helium [Abel & Haehnelt, 1999, McQuinn et al., 2009, Meiksin & Tittley, 2012] reveal that the heat injection and the resulting temperature evolution of the IGM depends on the details of how and when reionization occurred. There is evidence that the thermal vestiges of H I reionization heating may persist until as late as $z \sim 4-5$, and thus be observable in the Lyα forest [Cen et al., 2009, Furlanetto & Oh, 2009, Hui & Haiman, 2003], whereas for HeII reionization at $z \sim 3$, the Lyα forest is observable over the full duration of the phase transition. Finally, other processes could inject heat into the IGM and impact its thermal state, such as the large-scale structure shocks which eventually produce the Warm Hot Intergalactic Medium [WHIM;e.g. Cen & Ostriker, 1999, Davé et al., 2001, 1999], heating from galactic outflows [Cen & Ostriker, 2006, Kollmeier et al., 2006], photoelectric heating of dust grains [Inoue & Kamaya, 2003, Nath et al., 1999], cosmic-ray heating [Nath & Biermann, 1993], Compton-heating from the hard X-ray background [Madau & Efstathiou, 1999], X-ray preheating [Ricotti et al., 2005, Tanaka et al., 2012a], or blazar heating [Broderick et al., 2012, Chang et al., 2012, Pfrommer et al., 2012, Puchwein et al., 2012]. Precise constraints on the thermal state of the IGM would help determine the relative importance of photoheating from reionization and these more exotic mechanisms.

Despite all the successes of our current model of the IGM, precise constraints on its thermal state and concomitant constraints on reionization (and other exotic heat sources) remain elusive. Attempts to characterize the IGM thermal state from Lyα forest measurements have a long history. In the simplest picture, the gas in the IGM obeys a power law temperature-density relation $T = T_0(\rho/\bar{\rho})^{\gamma-1}$, which arises from the balance between photoionization heating, and cooling due to adiabatic expansion [Hui & Gnedin, 1997]. The standard approach has been to compare measurements of various statistics of the Lyα forest to cosmological hydrodynamical simulations. Leveraging the dependence of these statistics on the underlying temperature-density relation, its slope and amplitude (T_0, γ) parameters can be constrained. To this end a wide variety of statistics have been employed, such as the power spectrum [Viel et al., 2009, Zaldarriaga et al., 2001] or analogous statistics quantifying the small-scale power like wavelets [Garzilli et al., 2012, Lidz et al., 2009, Theuns et al., 2002b] or the curvature [Becker et al., 2011]. The flux PDF

[Bolton et al., 2008, Calura et al., 2012, Garzilli et al., 2012, Kim et al., 2007, McDonald et al., 2000] and the shape of the b-parameter distribution [Bryan & Machacek, 2000, Haehnelt & Steinmetz, 1998, McDonald et al., 2001, Ricotti et al., 2000, Rudie et al., 2012, Schaye et al., 2000, Theuns et al., 2000, 2002a] have also been considered. Multiple statistics have also been combined such as the PDF and wavelets [Garzilli et al., 2012], or PDF and power spectrum [Viel et al., 2009]. Overall, the results of such comparisons are rather puzzling. First, the IGM appears to be generally too hot, both at low ($z \sim 2$) and high ($z \sim 4$) redshift [Hui & Haiman, 2003]. In particular, the high inferred temperatures at $z \sim 4$ [e.g. Lidz et al., 2009, McDonald et al., 2001, Schaye et al., 2000, Theuns et al., 2002b, Zaldarriaga et al., 2001] suggest that HeII was reionized at still higher redshift $z > 4$ [Hui & Haiman, 2003], possibly conflicting with the late $z \sim 2.7$ reionization of HeII observed in HST/COS spectra [Furlanetto & Dixon, 2010, Shull et al., 2010, Syphers et al., 2012, Worseck et al., 2011, Worseck et al. 2013, in preparation]. Second, Bolton et al. [2008] considered the PDF of high-resolution quasar spectra and concluded that, at $z \simeq 3$ the slope of the temperature-density relation γ is either close to isothermal ($\gamma = 1$) or even inverted ($\gamma < 1$), suggesting "that the voids in the IGM may be significantly hotter and the thermal state of the low-density IGM may be substantially more complex than is usually assumed." Although this result is corroborated by additional work employing different statistics/methodologies [Calura et al., 2012, Garzilli et al., 2012, Viel et al., 2009, but see Lee et al. 2012], radiative transfer simulations of HeII reionization cannot produce an isothermal or inverted slope, unless a population other than quasars reionized HeII [Bolton et al., 2004, McQuinn et al., 2009, Meiksin & Tittley, 2012] , which would fly in the face of conventional wisdom. To summarize, despite nearly a decade of theoretical and observational work, published measurements of the thermal state of the IGM are still highly confusing, and concomitant constraints on reionization scenarios are thus hardly compelling.

Fortunately, there is another important record of the thermal history of the Universe: the Jeans pressure smoothing scale. Although baryons in the IGM trace dark matter fluctuations on large Mpc scales, on smaller scales $\lesssim 100$ kpc, gas is pressure supported against gravitational collapse by its finite temperature. Analogous to the classic Jeans argument, baryonic fluctuations are suppressed relative to the pressureless dark matter (which can collapse), and thus small-scale power is 'filtered' from the IGM [Gnedin & Hui, 1998], which explains why it is sometimes referred to as the *filtering scale*. Classically the *comoving* Jeans scale is defined as $\lambda_J^0 = \sqrt{\pi c_s^2/G\rho}(1+z)$, but in reality the amount of Jeans filtering is sensitive to both the instantaneous pressure and hence temperature of the IGM, *as well as the temperature of the IGM in the past.* This arises because fluctuations at earlier times expanded or failed to collapse depending on the IGM temperature at that epoch. Thus the Jeans scale reflects the competition between

gravity and pressure integrated over the Universe's history, and cannot be expressed as a mere deterministic function of the instantaneous thermal state. Heuristically, this can be understood because reionization heating is expected to occur on the reionization timescales of several hundreds of Myr, whereas the baryons respond to this heating on the sound-crossing timescale $\lambda_J^0/[c_s(1+z)] \sim (G\rho)^{-1/2}$, which at mean density is comparable to the Hubble time t_H.

Gnedin & Hui [1998] considered the behavior of the Jeans smoothing in linear theory, and derived an analytical expression for the filtering scale λ_J as a function of thermal history

$$\begin{aligned}\lambda_J^2(t) = \frac{1}{D_+(t)} \int_0^t dt' a^2(t') (\lambda_J^0(t'))^2 \times \\ (\ddot{D}_+(t') + 2H(t')\dot{D}_+(t')) \int_{t'}^t \frac{dt''}{a^2(t'')},\end{aligned} \tag{1.1}$$

where $D_+(t)$ is the linear growth function at time t, $a(t)$ is the scale factor, and $H(t)$ the Hubble expansion rate. Although this simple linear approximation provides intuition about the Jeans scale and its evolution, Fourier modes with wavelength comparable to the Jeans scale are already highly nonlinear at $z \sim 3$, and hence this simple linear pictures breaks down due to nonlinear mode-mode coupling effects. Thus given that we do not know the thermal history of the Universe, that we expect significant heat injection from HeII reionization at $z \sim 3-4$ concurrent with the epoch at which we observe the IGM, and that IGM modes comparable to the Jeans scale actually respond non-linearly to this unknown heating, the true relationship between the Jeans scale and the temperature-density relation at a given epoch should be regarded as highly uncertain.

Besides providing a thermal record of the IGM, the small-scale structure of baryons, as quantified by the Jeans scale, is a fundamental ingredient in models of reionization and galaxy formation. A critical quantity in models of cosmic reionization is the clumping factor of the IGM $C = \langle n_H^2 \rangle / \bar{n}_H^2$ [e.g. Emberson et al., 2013, Haardt & Madau, 2012, Madau et al., 1999, McQuinn et al., 2011, Miralda-Escudé et al., 2000, Pawlik et al., 2009], because it determines the average number of recombinations per atom, or equivalently the total number of UV photons needed to keep the IGM ionized. The clumping and the Jeans scale are directly related. Specifically,

$$C = 1 + \sigma_{\rm IGM}^2 \equiv 1 + \int d\ln k \, \frac{k^3 P_{\rm IGM}(k)}{2\pi^2}, \tag{1.2}$$

where $\sigma_{\rm IGM}^2$ is the variance of the IGM density, and $P_{\rm IGM}(k)$ is the 3D power spectrum of the baryons in the IGM. Given the shape of $P_{\rm IGM}(k)$, the integral above is dominated by contributions from small-scales (high-k), and most important is the Jeans cutoff

λ_J, which determines the maximum k-mode $k_{\rm J} \sim 1/\lambda_{\rm J}$ contributing. The small-scale structure of the IGM strongly influences the propagation of cosmological ionization fronts during reionization [Iliev et al., 2005]. Furthermore, several numerical studies have revealed that the hydrodynamic response of the baryons in the IGM to impulsive reionization heating is significant [e.g. Ciardi & Salvaterra, 2007, Gnedin, 2000a, Haiman et al., 2001, Kuhlen & Madau, 2005, Pawlik et al., 2009], indicating that a full treatment of the interplay between IGM small-scale structure and reionization history probably requires coupled radiative transfer hydrodynamical simulations.

Reionization heating also evaporates the baryons from low-mass halos or prevents gas from collapsing in them altogether [e.g. Barkana & Loeb, 1999, Dijkstra et al., 2004], an effect typically modeled via a critical mass, below which galaxies cannot form [Benson et al., 2002a,b, Bullock et al., 2000, Gnedin, 2000b, Kulkarni & Choudhury, 2011, Somerville, 2002]. Gnedin [2000b] used hydrodynamical simulations to show that this scale is well approximated by the *filtering mass*, which is the mass-scale corresponding to the Jeans filtering length, i.e. $M_F(z) = 4\pi\bar{\rho}\lambda_J^3/3$ [see also Hoeft et al., 2006, Okamoto et al., 2008]. Finally, because the Jeans scale has memory of the thermal events in the IGM (see eqn. 1.1), its value at later times can potentially constrain models of early IGM preheating. In this scenario, heat is globally injected into the IGM at high-redshift $z \sim 5-15$ from blast-waves produced by outflows from proto-galaxies or miniquasars [Benson & Madau, 2003, Cen & Bryan, 2001, Madau, 2000, Madau et al., 2001, Scannapieco et al., 2002, Scannapieco & Oh, 2004, Theuns et al., 2001, Voit, 1996] X-ray radiation from early miniquasars [Parsons et al., 2013, Tanaka et al., 2012b], which sets an entropy floor in the IGM and the raises filtering mass scale inhibiting the formation of early galaxies.

A rough estimate of the filtering scale at $z = 3$ can be obtained from eqn. (1.1) and the following simplified assumptions: the temperature at $z = 3$ is $T(z = 3) \approx 15000\,{\rm K}$ as suggested by measurements [e.g. Lidz et al., 2009, Ricotti et al., 2000, Schaye et al., 2000, Zaldarriaga et al., 2001], temperature evolves as $T \propto 1+z$, the typical overdensity probed by the $z = 3$ Lyα forest is $\delta \sim 2$ [Becker et al., 2011]. One then obtains $\lambda_J(z = 3) \approx 340$ kpc (comoving), smaller than the classical or instantaneous Jeans scale λ_J^0 by a factor of ~ 3. This distance maps to a velocity interval $v_J = Ha\lambda_J \approx 26\,{\rm km\ s^{-1}}$ along the line of sight due to Hubble expansion. Thermal Doppler broadening gives rise to a cutoff in the longitudinal power spectrum, which occurs at a comparable velocity $v_{\rm th} \approx 11.3\,{\rm km\ s^{-1}}$, for gas heated to the same temperature. The similarity of the characteristic scale of 3D Jeans pressure smoothing and the 1D thermal Doppler smoothing suggests that disentangling the two effects will be challenging given purely longitudinal observations of the Lyα forest, as confirmed by Peeples et al. [2009a], who considered the relative impact of thermal broadening and pressure smoothing on various

statistics applied to longitudinal Lyα forest spectra. Previous work that has aimed to measure thermal parameters such as T_0 and γ from Lyα forest spectra, have largely ignored the degeneracy of the Jeans scale with these thermal parameters. The standard approach has been to assume values of the Jeans scale from a hydrodynamical simulation [e.g. Becker et al., 2011, Lidz et al., 2009, Viel et al., 2009], which as per the discussion above, is equivalent to assuming perfect knowledge of the IGM thermal history. Because of the degeneracy with the Jeans scale, it is thus likely that previous measurements of the thermal parameters T_0 and γ are significantly biased, and their error bars significantly underestimated, if indeed Jeans scale takes on values different from those assumed (but see Zaldarriaga et al. 2001 who marginalized over the Jeans scale, and Becker et al. 2011 who also considered its impact). We will investigate such degeneracies in detail in chapter 2 with respect to power-spectra, and we consider degeneracies for a broader range of IGM statistics in a future work (A.Rorai et al., in preparation).

The Jeans filtering scale can be directly measured using close quasar pair sightlines which have comparable transverse separations $r_\perp \lesssim 300\,\mathrm{kpc}$ (comoving; $\Delta\theta \lesssim 40''$ at $z = 3$). The observable signature of Jeans smoothing is increasingly coherent absorption between spectra at progressively smaller pair separations resolving it [Peeples et al., 2009b]. The idea of using pairs to constrain the small scale structure of the IGM is not new. However, all previous measurements have either focused on lensed quasars, which probe extremely small transverse distances $r_\perp \sim 1\,\mathrm{kpc} \ll \lambda_J$ [e.g. McGill, 1990, Petry et al., 1998, Rauch et al., 2001, Smette et al., 1995, Young et al., 1981] such that the Lyα forest is essentially perfectly coherent, or real physical quasar pairs with $r_\perp \sim 1$ Mpc $\gg \lambda_J$ [D'Odorico et al., 2006] far too large to place useful constraints on the Jeans scale. Observationally, the breakthrough enabling a measurement of the Jeans scale is the discovery of a large number of close quasar pairs [Hennawi, 2004, Hennawi et al., 2009, 2006b, Myers et al., 2008] with $\sim 100\,\mathrm{kpc}$ separations. By applying machine learning techniques [Bovy et al., 2011, 2012, Richards et al., 2004] to the Sloan Digital Sky Survey [SDSS; York et al., 2000] imaging, a sample of ~ 300 close $r_\perp < 700\,\mathrm{kpc}$ quasar pairs at $1.6 < z \lesssim 4.3$[1] has been uncovered [Hennawi, 2004, Hennawi et al., 2009, 2006b].

In this paper we introduce a new method which enabled the first determination of the Jeans scale from a dataset of close quasar pair. We explicitly consider degeneracies between the canonical thermal parameters T_0 and γ, and the Jeans scale λ_J, which have been heretofore largely ignored. To this end, we use an approximate model of the Lyα forest based on dark matter only simulations, allowing us to independently vary all thermal parameters and simulate a large parameter space. The structure of the thesis is as follows: we describe how we compute the Lyα forest flux transmission from dark

[1] The lower redshift limit is corresponds to Lyα forest absorption being above the atmospheric cutoff.

matter simulations, and our parametrization of the thermal state of the IGM in section chapter 2. We focus in particular on the degeneracies between thermal parameters which result when only longitudinal observations are available, and how the additional transverse information provided by quasar pairs can break them. In chapter 3 we introduce our new method to quantify absorption coherence using the difference in phase between homologous longitudinal Fourier modes of each member of a quasar pair. We present a Bayesian likelihood formalism that uses the phase angle probability distributions to determine the Jeans scale, and we conduct a Markov Chain Monte Carlo (MCMC) analysis to determine the resulting precision on T_0, γ, and λ_J expected for realistic datasets, explore parameter degeneracies, and study the impact of noise and systematic errors.

The sample of observed pairs and the treatment of noise, resolution and contaminants are described in chapter 4, and the results obtained from the fully-calibrated phase difference analysis are shown in chapter 5. We also test the robustness of these results against a series of possible sources of bias. Chapter 6 addresses the problem of the physical interpretation of the Jeans scale measurement, using a set of hydrodynamic simulations, and illustrates a preliminary comparison of our estimate with the prediction of the standard model of the IGM on λ_J. We conclude and summarize in § 7.

Chapter 2

Parametric Study of the Intergalactic Medium

Our goal is to quantitatively assess the sensitivity of the transverse coherence in quasar pairs to the small-scale physics of the IGM, and to understand if the velocity-space degeneracy between thermal broadening and pressure support could be broken. To do this, we implement a machinery to rapidly predict Lyα-forest statistics in the space of parameters that describe the thermal state of the IGM. This machinery is based on two main components: a grid of thermal models of the IGM that sample the parameter space and a fast and flexible interpolation algorithm.

The thermal models are based on a Nbody dark-matter simulation, assuming that baryons trace dark matter and approximating the effect of pressure as a convolution with a *smoothing kernel* (see § 2.1.2). The width of this kernel defines in our model the Jeans filtering scale λ_J. The temperature is obtained by assuming a deterministic temperature-density relationship $T = T_0(1+\delta)^{\gamma-1}$. The triple $\{T_0, \gamma, \lambda_J\}$ defines the parameter space where the models reside.

A grid of models in this space constitutes the "training grid" for the *emulator*(§ 2.2), an algorithm based on principal component decomposition and Gaussian-processes interpolation that allows to efficiently predict the Lyα-forest statistics at any value of T_0, γ and λ_J.

We conclude the chapter showing an application of this emulator to the line-of-sight power spectrum and the cross power spectrum (§ 2.3), showing explicitly the degeneracy between the thermal parameters.

Here and in the next chapter we use the ΛCDM cosmological model with the parameters $\Omega_m = 0.28, \Omega_\Lambda = 0.72, h = 0.70, n = 0.96, \sigma_8 = 0.82$. All distances quoted are in comoving kpc.

2.1 Simulation

2.1.1 Dark Matter Simulation

Our model of the Lyα forest is based on a Nbody dark matter only simulation. In this scheme, the dark matter simulation provides the dark matter density and velocity field [Croft et al., 1998, Meiksin & White, 2001], and the gas density and temperature are computed using simple scaling relations motivated by the results of full hydrodynamical simulations [Gnedin et al., 2003, Gnedin & Hui, 1998, Hui & Gnedin, 1997]. Our objective is then to explore the sensitivity with which close quasar pairs can be used to constrain the thermal parameters defining these scaling relations, and in particular the Jeans scale. To this end, we require a dense sampling of the thermal parameter space, which is computationally feasible with our semi-analytical method applied to a dark matter simulation snapshot, whereas it would be extremely challenging to simulate such a dense grid with full hydrodynamical simulations. We do not model the redshift evolution of the IGM, nor do we consider the effect of uncertainties on the cosmological parameters, as they are constrained by various large-scale structure and CMB measurements to much higher precision than the thermal parameters governing the IGM.

We used an updated version version of the TreePM code described in White [2002] to evolve 1500^3 equal mass ($3\times10^6\, h^{-1} M_\odot$) particles in a periodic cube of side length $L_{\rm box} = 50\, h^{-1}$Mpc with a Plummer equivalent smoothing of $1.2\, h^{-1}$kpc. The initial conditions were generated by displacing particles from a regular grid using second order Lagrangian perturbation theory at $z = 150$. This TreePM code has been compared to a number of other codes and has been shown to perform well for such simulations [Heitmann et al., 2008]. Recently the code has been modified to use a hybrid MPI+OpenMP approach which is particularly efficient for modern clusters.

In this and in the next chapter, we analyze the snapshot at $z = 3$

2.1.2 Description of the Intergalactic Medium

The baryon density field is obtained by smoothing the dark matter distribution; this smoothing mimics the effect of the Jeans pressure smoothing. For any given thermal

model, we adopt a constant filtering scale λ_J, rather than computing it as a function of the temperature, and this value is allowed to vary as a free parameter (see discussion below). The dark matter distribution is convolved with a window function $W_{\rm IGM}$, which, in Fourier space, has the effect of quenching high-k modes

$$\delta_{\rm IGM}(\vec{k}) = W_{\rm IGM}(\vec{k}, \lambda_J)\delta_{\rm DM}(\vec{k}) \tag{2.1}$$

For example a Gaussian kernel with $\sigma = \lambda_J$, $W_{\rm IGM}(k) = \exp(-k^2\lambda_J^2/2)$, would truncates the 3D power spectrum at $k \sim 1/\lambda_J$.

Because we smooth the dark matter particle distribution in real-space, it is more convenient to adopt a function with a finite-support

$$\delta_{\rm IGM}(x) \propto \sum_i m_i K(|x - x_i|, R_J) \tag{2.2}$$

where m_i and x_i are the mass and position of the particle i, $K(r)$ is the kernel, and R_J the smoothing parameter which sets the Jeans scale. We adopt the followoing cubic spline kernel

$$K(r, R_J) = \frac{8}{\pi R_J^3} \begin{cases} 1 - 6\left(\frac{r}{R_J}\right)^2 + 6\left(\frac{r}{R_J}\right)^3 & \frac{r}{R_J} \leq \frac{1}{2} \\ 2\left(1 - \frac{r}{R_J}\right)^3 & \frac{1}{2} < \frac{r}{R_J} \leq 1 \\ 0 & \frac{r}{R_J} > 1 \end{cases} . \tag{2.3}$$

In the central regions the shape of $K(r)$ very closely resembles a Gaussian with $\sigma \sim R_J/3.25$, and we will henceforth take this $R_J/3.25$ to be our definition of λ_J, which we will alternatively refer to as the 'Jeans scale' or the 'filtering scale'. The analogous smoothing procedure is also applied to the particle velocities; however, note that the velocity field has very little small-scale power, and so the velocity distribution is essentially unaffected by this pressure smoothing operation. As we discuss further in Appendix A, the mean inter-particle separation of our simulation cube $\delta l = L_{\rm box}/N_{\rm p}^{1/3}$ sets the minimum Jean smoothing that we can resolve with our dark matter simulation, hence we can safely model values of $\lambda_J > 42$kpc.

At the densities typically probed by the Lyα forest, the IGM is governed by relatively simple physics. Most of the gas has never been shock heated, is optically thin to ionizing radiation, and can be considered to be in ionization equilibrium with a uniform UV background. Under these conditions, the competition between photoionization heating and adiabatic expansion cooling gives rise to a tight relation between temperature and

density which is well approximated by a power law [Hui & Gnedin, 1997],

$$T(\delta) = T_0(1+\delta)^{\gamma-1} \tag{2.4}$$

where T_0, the temperature at the mean density, and γ, the slope of the temperature-density relation, both depend on the thermal history of the gas. We thus follow the standard approach, and parametrize the thermal state of the IGM in this way. Typical values for T_0 are on the order of 10^4 K, while γ is expected to be around unity, and asymptotically approach the value of $\gamma_\infty = 1.6$, if there is no other heat injection besides (optically thin) photoionzation heating. Recent work suggests that an inverted temperature-density relation $\gamma < 1$ provides a better match to the flux probability distribution of the Lyα forest [Bolton et al., 2008], but the robustness of this measurement has been debated [Lee, 2012].

The optical depth for Lyα absorption is proportional to the density of neutral hydrogen n_{HI}, which, if the gas is highly ionized ($x_{HI} \ll 1$) and in photoionization equilibrium, can be calculated as [Gunn & Peterson, 1965]

$$n_{HI} = \alpha(T) n_H^2 / \Gamma \tag{2.5}$$

where Γ is the photoionization rate due to a uniform metagalactic ultraviolet background (UVB), and $\alpha(T)$ is the recombination coefficient which scales as $T^{-0.7}$ at typical IGM temperatures. These approximations result in a power law relation between Lyα optical depth and overdensity often referred as the fluctuating Gunn-Petersonn approximation (FGPA) $\tau \propto (1+\delta)^{2-0.7(\gamma-1)}$, which does not include the effect of peculiar motions and thermal broadening. We compute the observed optical depth in redshift-space via the following convolution of the real-space optical depth

$$\tau(v) = \int_{-\infty}^{\infty} \tau(x)\Phi(Hax + v_{p,\parallel}(x) - v, b(x))dx, \tag{2.6}$$

where Hax is the real-space position in velocity units, $v_{p,\parallel}(x)$ is the longitudinal component of the peculiar velocity of the IGM at location x, and Φ is the normalized Voigt profile (which we approximate with a Gaussian) characterized by the thermal width $b = \sqrt{2K_BT/mc^2}$, where we compute the temperature from the baryon density via the temperature-density relation (see eqn. 2.4). The observed flux transmission is then given by $F(v) = e^{-\tau(v)}$.

We apply the aforementioned recipe to 2×100^2 lines-of-sight (*skewers*) running parallel to the box axes, to generate the spectra of 100^2 quasar pairs, and we repeat this procedure for 500 different choices of the parameter set (T_0, γ, λ_J). Half of the spectra (the first member of each pair) are positioned on a regular grid in the $y-z$ plane,

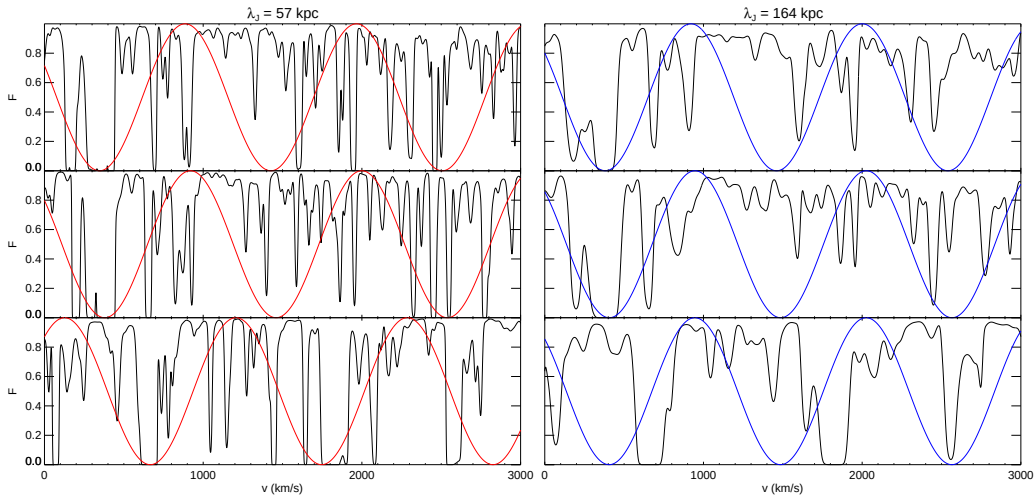

FIGURE 2.1: An example of three simulated spectra. The left and the right panels represent the same spectra in the simulation calculated for two models with different Jeans smoothing length λ_J. The middle and the lower panel represent two spectra respectively at separation 0.5 Mpc and 1 Mpc from the top one. The coloured sine curves track homologous Fourier modes in each spectrum, with rescaled mean and amplitude to fit the range $[0, 1]$. The wave shifts provide a graphical visualization of phase differences, which we will use to quantify spectral coherence and probe the Jeans scale (see chapter 3). The right panels suggest that a larger λ_J results in greater spectral coherence and generally smaller phase differences between neighboring sightlines.

in order to distribute them evenly in space. Subsequently, a companion is assigned to each of them, and our choice for the distribution of radial distances warrants further discussion. Our goal is to statistically characterize the coherence of pairs of spectra as a function of impact parameter, and near the Jeans scale this coherence varies rapidly with pair separation. Hence computing statistics in bins of transverse separation is undesirable, because it can lead to subtle biases in our parameter determinations if the bins are too broad. To circumvent these difficulties, we focus our entire analysis on 30 linearly-spaced discrete pair separations between 0 and 714 kpc. For each of the 100^2 lines-of-sight on the regular grid, a companion sightline is chosen at one of these discrete radial separations, where the azimuthal angle is drawn from a uniform distribution.

We follow the standard approach, and treat the metagalactic photoionization rate Γ as a free parameter, whose value is fixed *a posteriori* by requiring the mean flux of our Lyα skewers $\langle \exp(-\tau) \rangle$ to match the measured values from Faucher-Giguere et al. [2007]. This amounts to a simple constant re-scaling of the optical depth. The value of the mean flux at $z = 3$ is taken to be fixed, and thus assumed to be known with infinite precision. This is justified, because in practice, the relative measurement errors on the mean flux are very small in comparison to uncertainties of the thermal parameters we wish to study. In a future work, we conduct a full parameter study using other Lyα forest statistics, and explore the effect of uncertainties of the mean flux (A.Rorai et al. 2013, in preparation). Examples of our spectra are shown in Figure 2.1.

To summarize, our models of the Lyα forest are uniquely described by the three parameters (T_0, γ, λ_J), and we reiterate that these three parameters are considered to be independent. In particular the Jeans scale is not related to the instantaneous temperature at mean density T_0. Although this may at first appear unphysical, it is motivated by the fact that λ_J depends non-linearly on the entire thermal history of the IGM (see eqn. 1.1), and both this dependence and the thermal history are not well understood, as discussed in the introduction. Allowing λ_J to vary independently is the most straightforward parametrization of our ignorance. However, improvements in our theoretical understanding of the relationship between λ_J and the thermal history of the IGM (T_0,γ) could inform more intelligent parametrizations. Furthermore, inter-dependencies between thermal parameters can also be trivially included into our Bayesian methodology for estimating the Jeans scale as conditional priors, e.g. $P(\lambda_J, T_0)$, in the parameter space.

2.2 emulator

Our goal is to define an algorithm to calculate $\zeta(k, r_\perp|T_0, \gamma, \lambda_J)$ as a function of the thermal parameters, interpolating from the values determined on a fixed grid. As we will also compare Jeans scale constraints from the phase angle PDF (eqn. 3.13), to those obtained from other statistics, such as the longitudinal power $P(k)$ and cross-power $\pi(k, r_\perp)$ (see § 3.2), we also need to be able to smoothly interpolate these functions as well. To achieve this, we follow the approach of the 'Cosmic Calibration Framework' (CCF) to provide an accurate prediction scheme for cosmological observables [Habib et al., 2007, Heitmann et al., 2006]. The aim of the CCF is to build *emulators* which act as very fast – essentially instantaneous – prediction tools for large scale structure observables such as the nonlinear power spectrum [Heitmann et al., 2009, 2010, Lawrence et al., 2010], or the concentration-mass relation [Kwan et al., 2012]. Three essential steps form the basis of emulation. First, one devises a sophisticated space-filling sampling scheme that provides an optimal sampling strategy for the cosmological parameter space being studied. Second, a principle component analysis (PCA) is conducted on the measurements from the simulations to compress the data onto a minimal set of basis functions that can be easily interpolated. Finally, Gaussian process modeling is used to interpolate these basis functions from the locations of the space filling grid onto any value in parameter space. A detailed description of our IGM emulator will be described in a forthcoming paper (A.Rorai et al., in preparation). Below we briefly summarize the key aspects.

2.2.1 Models

Whereas CCF uses more sophisticated space filling Latin Hypercube sampling schemes [e.g. Heitmann et al., 2009], we adopt a simpler approach motivated by the shape of the IGM statistics we are trying to emulate, which change rapidly at scales comparable to either the Jeans or thermal smoothing scale. We opt for an irregular scattered grid which fills subspaces more effectively than a cubic lattice. We consider parameter values over the domain $\{(T_0, \gamma, \lambda_J) : T_0 \in [5000, 40000]\,\mathrm{K};\ \gamma \in [0.5, 2];\ \lambda_J \in [43, 572]\,\mathrm{kpc}\}$. The lower limit of 43 kpc for the Jeans scale is chosen because this is about the smallest value we can resolve with our simulation (see Appendix A), while the upper limit of 572 kpc is a conservative constraint deduced from the longitudinal power spectrum: a filtering scale greater than this value would be inconsistent with the high$-k$ cutoff, regardless of the value of the temperature. The ranges considered for T_0 and γ are consistent with those typically considered in the literature and our expectations based on the physics governing the IGM. We sample the 3D thermal parameter space at 500 locations, where we consider a discrete set of 50 points in each dimension. A linear spacing of these points is adopted for γ, whereas we find it more appropriate to distribute T_0 and λ_J such that the scale of the cutoff of the power spectrum k_f is regularly spaced. Since $k_f \propto \lambda_J^{-1}$ for Jeans smoothing and $k_f \propto T_0^{-1/2}$ for thermal broadening, we choose regular intervals of these parameters after transforming $\lambda_J \to 1/\lambda_J$ and $T_0 \to 1/\sqrt{T_0}$. Each of the 50 values of the parameters is then repeated exactly 10 times in the 500-point grid, and we use 10 different random permutations of their indices to fill the space and to avoid repetition. For each thermal model in this grid, we generate 10,000 pairs of skewers at 30 linearly spaced discrete pair separations between 0 and 714 kpc.

2.2.2 PCA

We then use these skewers to compute the IGM statistics $\zeta(k, r_\perp)$, $P(k)$, and $\pi(k, r_\perp)$ for all k and $r_\perp$ for each thermal model. A PCA decomposition is then performed in order to compress the information present in each statistic and represent its variation with the thermal parameters using a handful of basis functions ϕ. A PCA is an orthogonal transformation that converts a family of correlated variables into a set of linearly uncorrelated combinations of principal components. The components are ordered by the variance along each basis dimension, thus relatively few of them are sufficient to describe the entire variation of a function in the space of interest, which is here the thermal parameter space. To provide a concrete example, the longitudinal power spectrum $P(k)$ is fully described by the values of the power in each k bin, but it is likely that some of these $P(k)$ values do not change significantly given certain combinations of thermal

parameters. The PCA determines basis functions of the $P(k)$ that best describe its variation with thermal parameters, enabling us to represent this complex dependence with an expansion onto just a few principal components

$$P(k|T_0, \gamma, \lambda_J) = \sum_i \omega_i(T_0, \gamma, \lambda_J)\Phi_i(k), \tag{2.7}$$

where $\{\Phi(k)\}$ are the basis of principal components, and $\{\omega\}$ are the corresponding coefficients which depend on the thermal parameters. The number of components for a given function is set by the maximum tolerable interpolation errors of the emulator, and these are in turn set by the size of the error bars on the statistic that one is attempting to model. We note that the number of PCA components we used to fully represent the functions $\zeta(k, r_\perp)$, $P(k)$, and $\pi(k, r_\perp)$ were 25, 15, and 25, respectively (phase distribution and cross power spectrum are 2D functions, so they need more components). We verified that adding further components did not change significantly our main results, indicating that we achieve convergence.

2.2.3 Gaussian Process Interpolation

Gaussian process interpolation is then used to interpolate these PCA coefficients $\omega_i(T_0, \gamma, \lambda_J)$ from the irregular distribution of points in our thermal grid to any location of interest in the parameter space. The only input for the Gaussian interpolation is the choice of *smoothing length*, which quantifies the degree of smoothness of each function along the direction of a given parameter in the space. We choose these smoothing lengths to be a multiple of the spacing of our parameter grid. The choice of these smoothing lengths is somewhat arbitrary, but we checked that the posterior distributions of thermal parameters (eqn. 3.13) inferred do not change in response to a reasonable variations of these smoothing lengths. A full description of the calibration and testing of the emulator is presented in an upcoming paper (Rorai et al., in prep).

2.3 Power Spectra and Their Degeneracies

Although many different statistics have been employed to isolate and constrain the thermal information contained in Lyα forest spectra, the flux probability density function (PDF; 1-point function) and the flux power spectrum or auto-correlation function (2-point function), are among the most common[e.g. Kim et al., 2007, McDonald et al., 2000, Viel et al., 2009, Zaldarriaga et al., 2001]. But because the Lyα transmission F is significantly non-Gaussian, significant information is also contained in higher-order statistics. For example wavelet decompositions, which contains a hybrid of real-space

and Fourier-space information, have been advocated for measuring spatial temperature fluctuations [Garzilli et al., 2012, Lidz et al., 2009, Zaldarriaga, 2002]. Several studies have focused on the on the b-parameter distribution to obtain constraints on thermal parameters [McDonald et al., 2001, Ricotti et al., 2000, Rudie et al., 2012, Schaye et al., 2000], and recently Becker et al. [2011] introduced a 'curvature' statistic as an alternative measure of spectral smoothness to the power spectrum.

As gas pressure acts to smooth the baryon density field in 3D, it is natural explore power spectra as a means to constrain the Jeans filtering scale. A major motivation for working in Fourier space, as opposed to the real-space auto-correlation function, is that it is much easier to deal with limited spectral resolution in Fourier space. The vast majority of close quasar pairs are too faint to be observed at echelle resolution FWHM $\simeq 5\,\mathrm{km\,s^{-1}}$ where the Lyα forest is completely resolved. Instead, spectral resolution has to be explicitly taken into account. But to a very good approximation the smoothing caused by limited spectral resolution simply low-pass filters the flux, and thus the shape of the flux power spectrum is unchanged for k-modes less than the spectral resolution cutoff $k_{\rm res}$. Thus by working in k-space, one can simply ignore modes $k \gtrsim k_{\rm res}$ and thus obviate the need to precisely model the spectral resolution, which can be challenging for slit-spectra. Finally, another advantage to k-space is that, because fluctuations in the IGM are only mildly non-linear, some of the desirable features of Gaussian random fields, such as the statistical independence of Fourier modes, are approximately retained, simplifying error analysis. In what follows we consider the impact of Jeans smoothing on longitudinal power spectrum, as well as the simplest 2-point function that can be computed from quasar pairs, the cross-power spectrum.

2.3.1 The Longitudinal Power Spectrum

It is well known that the shape of the longitudinal power spectrum, and the high-k thermal cutoff in particular, can be used constrain the T_0 and γ [Viel et al., 2009, Zaldarriaga et al., 2001]. This cutoff arises because thermal broadening smooths τ in redshift-space (e.g. eqn. 2.6). In contrast to this 1D smoothing, the Jeans filtering smooths the IGM in 3D, and it is exactly this confluence between 1D and 3D smoothing that we want to understand [see also Peeples et al., 2009a,b]. We consider the quantity $\delta F(v) = (F - \bar{F})/\bar{F}$, where $\bar{F}$ is the mean transmitted flux, and compute the power spectrum according to

$$P(k) = \langle |\delta\tilde{F}(k)|^2 \rangle, \tag{2.8}$$

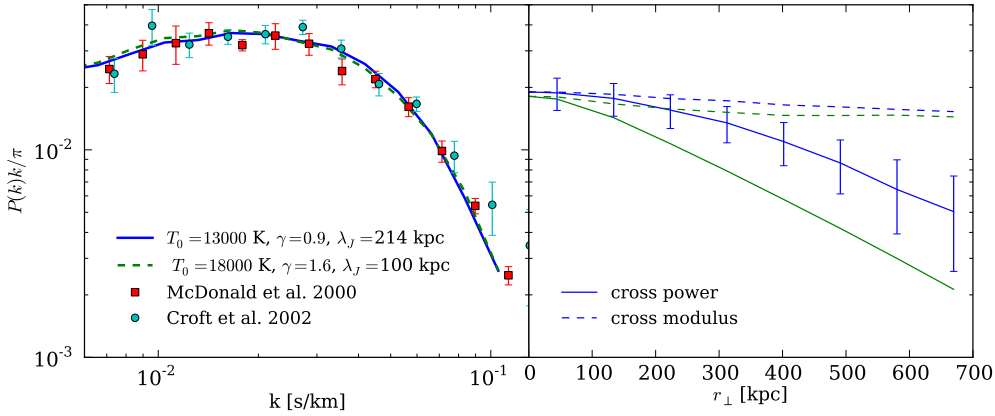

FIGURE 2.2: *Left panel:* The 1D dimensionless power spectrum of the Lyα forest at $z = 3$. In our large grid of thermal models, we can identify two very different parameter combinations, represented by the solid (blue) and dashed (green) curves, which provide an equally good fit to the longitudinal power spectrum measurements from McDonald et al. [2000] (red squares) and Croft et al. [2002] (cyan circles), illustrating the strong degeneracies between these parameters (T_0,γ,λ_J). In light of these degeneracies, it is clear that it would be extremely challenging to constrain these parameters with the longitudinal power alone. *Right panel:* The dimensionless cross power spectrum $\pi(k;r_\perp)k/\pi$ (solid line) at $k \approx 0.05$ s/km from our simulated skewers, as a function of $r_\perp$ for the same two thermal models shown at left, with error bars estimated from a sample of 20 pairs. The degeneracy afflicting the 1D power is broken using the new information provided by close quasar pairs, because the different Jeans scales result in differing amounts of transverse spectral coherence, providing much better prospects for measuring λ_J. We also show the cross modulus $\langle\rho_1(k)\rho_2(k)\rangle k/\pi$ (dashed lines) for the same two models, which show flat variation with $r_\perp$, and a very weak dependence on the Jeans scale. Most of the information about the 3D Jeans smoothing resides not in the moduli, but rather in the phase differences between homologous modes (see discussion in § 3.1.3).

where $\delta\tilde{F}(k)$ denotes the Fourier transform of δF for longitudinal wavenumber k, and angular brackets denote an suitable ensemble average (i.e. over our full sample of spectra).

In Figure 2.2 we compare two thermal models in our thermal parameter grid to measurements of the longitudinal power spectrum of the Lyα forest at $z \simeq 3$ [Croft et al., 2002, McDonald et al., 2000]. The blue (solid) curve has a large Jeans scale $\lambda_J = 214\,\mathrm{kpc}$, a cooler IGM $T_0 = 13,000\,\mathrm{K}$, and a nearly isothermal temperature-density relation $\gamma = 0.9$, which is mildly inverted such that voids are hotter than overdensities. Such isothermal or even inverted equations of state could arise at $z \sim 3$ from He II reionization heating [McQuinn et al., 2009, Tittley & Meiksin, 2007b], and recent analyses of the flux PDF [Bolton et al., 2008] as well joint analysis of PDF and power-spectrum [Calura et al., 2012, Garzilli et al., 2012, Viel et al., 2009] have argued for inverted or nearly isothermal values of γ. The green (dashed) curves have a smaller Jeans scale $\lambda_J = 100\,\mathrm{kpc}$, a hotter IGM $T_0 = 18,000\,\mathrm{K}$, and a steep $\gamma = 1.6$ temperature-density relation consistent with the asymptotic value if the IGM has not undergone significant

recent heating events [Hui & Gnedin, 1997, Hui & Haiman, 2003]. Thus with regards to the longitudinal power spectrum, the Jeans scale is clearly degenerate with the amplitude and slope (T_0, γ) of the temperature-density relation. One would clearly come to erroneous conclusions about the equation of state parameters (T_0,γ) from longitudinal power spectrum measurements, if the lack of knowledge of the Jeans scale is not marginalized out [see e.g. Zaldarriaga et al., 2001, for an example of this marginalization].

This degeneracy in the longitudinal power arises because the Jeans filtering smooths the power in 3D on a scale which project to a longitudinal velocity

$$v_J = \frac{H(z=3)}{1+3}\lambda_J \approx 26 \left(\frac{\lambda_J}{340 \text{ kpc}}\right) \text{ km s}^{-1}, \tag{2.9}$$

resulting in a cutoff of the power at $k_J \approx 0.04\,\text{s km}^{-1}$ (for the typical values assumed in the introduction[1]). The thermal Doppler broadening of Lyα absorption lines smooths the power in 1D, on a scale governed by the *b-parameter*

$$b = \sqrt{\frac{2k_B T}{\mu m_p}} \approx 15.7 \left(\frac{T}{1.5\times10^4 \text{ K}}\right)^{1/2} \text{ km s}^{-1}, \tag{2.10}$$

which results in an analogous cutoff at $k_{\rm th} = \sqrt{2}/b \approx 0.09\,\text{s km}^{-1}$ for a temperature of 15000 K. Above k_B is the Boltzmann constant, m_p the proton mass, and $\mu \approx 0.59$ is the mean molecular weight for a primordial, fully ionized gas. The fact that the two cutoff scales are comparable results in a strong degeneracy which is very challenging to disentangle with longitudinal observations alone. Similar degeneracies between the Jeans scale and (T_0,γ) exist if one considers wavelets, the curvature, the b-parameter distribution, and the flux PDF, which we explore in an upcoming study (Rorai et al. 2013, in prep). In the next section we show that this degeneracy between 3D and 1D smoothing can be broken by exploiting additional information in the transverse dimension provided by close quasar pairs.

2.3.2 Cross Power Spectrum

The foregoing discussion illustrates that the 3D (Jeans) and 1D (thermal broadening) smoothing are mixed in the longitudinal power spectrum, and ideally one would measure the full 3D power spectrum to break this degeneracy. For an isotropic random field the

[1] We caution that this estimate assumes a thermal history where $T \propto 1+z$, without considering the effect of HeII reionization. In that case the deduced value for the filtering scale λ_J would probably be smaller.

1D power spectrum $P(k)$ and the 3D power $P_{3D}(k)$ are related according to

$$P_{3D} = \frac{1}{2\pi}\frac{1}{k}\frac{dP(k)}{dk}. \tag{2.11}$$

However, in the Lyα forest redshift-space distortions and thermal broadening result in an anisotropies that render this expression invalid.

With close quasar pairs, transverse correlations measured across the beam contain information about the 3D power, and can thus thus disentangle the 3D and 1D smoothing. Consider for example the cross-power spectrum $\pi(k, r_\perp)$ of two spectra $\delta F_1(v)$ and $\delta F_2(v)$ separated by a transverse distance $r_\perp$

$$\pi(k; r_\perp) = \Re[\delta\tilde{F}_1^*(k)\delta\tilde{F}_2(k)]. \tag{2.12}$$

When $r_\perp \to 0$ then $\delta F_2 \to \delta F_1$ and the cross-power tends to the longitudinal power $P(k)$. The cross-power can be thought of as effectively a power spectrum in the longitudinal direction, and a correlation function in the transverse direction [see also Viel et al., 2002]. Alternatively stated, the cross power provides a transverse distance dependent correction to the longitudinal power $P(k)$, reducing it from its maximal value at 'zero lag' $r_\perp = 0$. This further implies that measuring the cross power of closely separated and thus highly coherent spectra amounts to, at some level, a somewhat redundant measurement of the longitudinal power which could be simply deduced from isolated spectra. In the next chapter, we will explain how to isolate the genuine 3D information provided by close quasar pairs using a statistic that is more optimal than the cross-power. Nevertheless, Figure 2.2 shows the cross-power spectrum for the two degenerate models discussed in the previous section, clearly illustrating that even the sub-optimal cross-power spectrum can break the strong degeneracies between thermal parameters that are present if one considers the longitudinal power alone.

Chapter 3

Phase Analysis of the Lyman-α Forest of Quasar Pairs

In the previous chapter I described the general method that we use to estimate the capability of a given Lyα-forest statistic of discriminating among different thermal models. Now we need to decide which statistic we want to apply to quasar pairs in order to extract the transverse coherence information. Our assessment of the ability of quasar pairs in pinpointing the Jeans scale will be strongly dependent on this choice.

The ideal statistic would have the property of being sensitive to the real-space coherence of density structure, while being independent on the velocity-space effect such as thermal broadening and redshift distortions due to peculiar velocities. In doing so, we will eliminate part of the information contained in the spectra which is intrinsically 1-dimensional. There are at least two good reasons to proceed in this way: the 1-d properties of the Lyαforest can be studied more effectively in spectra of individual QSOs at the same redshifts, which are more frequent and brighter than pairs; along the line of sight, real-space and velocity-space effects exhibit degeneracies which are difficult to treat. Moreover, an high sensitivity to redshift-space distortions would raise the requirements on our theoretical understanding and on the details of our model, challenging the capabilities of the simple models that we employ.

In this chapter I will explain how this is achieved by adopting the phase-difference statistic, whereas the use of the most obvious transverse statistic, i.e. the cross power or the cross correlation function, would have been ineffective.

3.1 A New Statistic: Phase Differences

Although the cross-power has the ability to break the degeneracy between 3D and 1D smoothing present in the longitudinal power, we demonstrate here that the cross-power (or equivalently the cross-correlation function) is however not optimal, and indeed the genuine 3D information is encapsulated in the *phase differences* between homologous Fourier modes.

3.1.1 Drawbacks of the Cross Power Spectrum

Let us write the 1D Fourier transform of the field δF as

$$\delta\tilde{F}(k) = \rho(k)e^{i\theta(k)} \tag{3.1}$$

where the complex Fourier coefficient is described by a modulus ρ and phase angle θ, both of which depend on k. Note that for any ensemble of spectra $P(k) = \langle\rho^2(k)\rangle$, hence the modulus $\rho(k)$ is a random draw from a distribution whose variance is given by the power spectrum. From eqn. (2.12), the cross-power of the two spectra $\delta F_1(v)$ and $\delta F_2(v)$ is then

$$\pi_{12}(k) = \rho_1(k)\rho_2(k)\cos(\theta_{12}(k)), \tag{3.2}$$

where $\theta_{12}(k) = \theta_1(k) - \theta_2(k)$ is the phase difference between the homologous $k-$modes. The distribution of the moduli ρ_1 and ρ_2 are also governed by $P(k)$, but at small impact parameter they are not statistically independent because of spatial correlations. Nevertheless, the moduli contain primarily information already encapsulated in the longitudinal power, and are thus affected by the same thermal parameter degeneracies that we described in the previous section. For the purpose of constraining the Jeans scale, we thus opt to ignore the moduli ρ_1 and ρ_2 altogether, in an attempt to isolate the genuine 3D information, increasing sensitivity to the Jeans scale, while minimizing the impact of thermal broadening, removing degeneracies with the temperature-density relation parameters (T_0,γ).

The foregoing points are clearly illustrated by the dashed curves in the right panel of Figure 2.2, which compares the quantity $\langle\rho_1(k)\rho_2(k)\rangle$ as a function of impact parameter $r_\perp$ for the same pair of thermal models discussed in § 2.3.1, which are degenerate with respect to the longitudinal power. The similarity of these two curves reflects the degeneracy of the longitudinal power for these two models, and one observes a flat trend with $r_\perp$ and a very weak dependence on the Jeans scale λ_J, substantiating our argument that the moduli contain primarily 1D information.

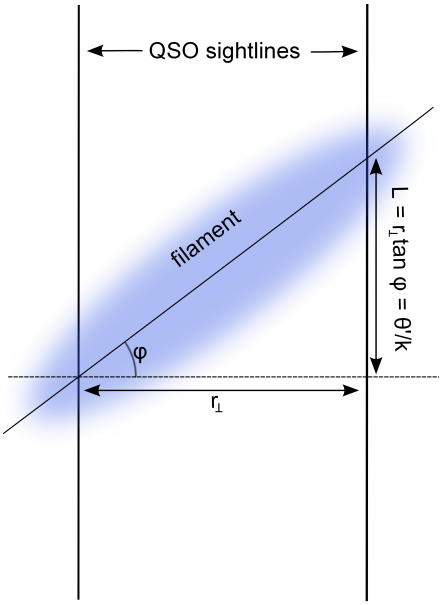

FIGURE 3.1: Schematic representation of the heuristic argument used to determine the phase difference distribution: phase are determined by density filaments crossing the lines of sight of two quasars. If the orientation of the filaments φ is isotropically distributed then θ', dependent on the longitudinal distance $L = r_\perp \tan\varphi$, follows a Cauchy distribution.

As the moduli contain minimal information about the 3D power, we are thus motivated to explore how the phase difference $\theta_{12}(k)$ can constrain the Jeans scale. In terms of Fourier coefficients, $\theta_{12}(k)$ can be written

$$\theta_{12}(k) = \arccos\left(\frac{\Re[\delta\tilde{F}_1^*(k)\delta\tilde{F}_2(k)]}{\sqrt{|\delta\tilde{F}_1(k)|^2|\delta\tilde{F}_2(k)|^2}}\right). \tag{3.3}$$

Note that because the phase difference is given by a ratio of Fourier modes, it is completely insensitive to the normalization of δF, and hence to quasar continuum fitting errors, provided that these errors do not add power on scales comparable to k. In the remainder of this section, we provide a statistical description of the distribution of phase differences and we explore the properties and dependencies of this distribution. To simplify notation we will omit the subscript and henceforth denote the phase difference as simply $\theta(k, r_\perp) = \theta_1(k) - \theta_2(k)$, where $r_\perp$ is the transverse distance between the two spectra $\delta F_1(v)$ and $\delta F_2(v)$.

3.1.2 An Analytical Form for the PDF of Phase Differences

The phase difference between homologous k-modes is a random variable in the domain $[-\pi, \pi]$, which for a given thermal model, depends on two quantities: the longitudinal mode in question k and the transverse separation $r_\perp$. One might advocate computing

the quantity $\langle \cos\theta(k, r_\perp)\rangle$ analogous to the cross-power (see eqn. 2.12), or the mean phase difference $\langle \theta(k, r_\perp)\rangle$, to quantify the coherence of quasar pair spectra. However, as we will see, the distribution of phase differences is not Gaussian, and hence is not fully described by its mean and variance. This approach would thus fail to exploit all the information encoded in its shape. Our goal is then to determine the functional form of the distribution of phase differences at any $(k, r_\perp)$, and relate this to the thermal parameters governing the IGM. This is a potentially daunting task, since it requires deriving a unique function in the 2-dimensional space $\theta(k, r_\perp)$ for any location in our 3-dimensional thermal parameter grid (T_0, γ, λ_J). Fortunately, we are able to reduce the complexity considerably by deriving a simple analytical form for the phase angle distribution.

We arrive at a this analytical form via a simple heuristic argument, whose logic is more intuitive in real space. Along the same lines, we focus initially on the IGM density distribution along 1D skewers, and then later demonstrate that the same form also applies to the Lyα flux transmission. Consider a filament of the cosmic web pierced by two quasar sightlines separated by $r_\perp$, and oriented at an angle φ relative to the transverse direction. A schematic representation is shown in Figure 3.1. This structure will result in two peaks in the density field along the two sightlines, separated by a longitudinal distance of $L = r_\perp \tan\varphi$. If we assume that the positions of these density maxima dictate the position of wave crests in Fourier space, the phase difference for a mode with wave number k can be written as $\theta' = kL = kr_\perp \tan\varphi$. We can derive the probability distribution of the phase difference by requiring that $p(\theta')d\theta' = p(\varphi)d\varphi$, and assuming that, by symmetry, φ is uniformly distributed. This implies that θ' follows the Cauchy-distribution

$$p(\theta') = \frac{1}{\epsilon\pi}\frac{1}{1 + (\theta'/\epsilon)^2}, \tag{3.4}$$

where ϵ parametrizes the distribution's concentration. As a final step, we need to redefine the angles such that they reside in the proper domain. Because $\tan\varphi$ spans the entire real line, so will θ'; however, for any integer n, all phases $\theta' + 2\pi n$ corresponding to distances $L + 2\pi n/k$ will map to identical values of θ, defined to be the phase difference in the domain $[-\pi, \pi]$. Redefining the domain, requires that we re-map our probabilities according to

$$P_{[-\pi,\pi]}(\theta) = \sum_{n\in\mathbb{Z}} p(\theta + 2\pi n), \tag{3.5}$$

a procedure known as 'wrapping' a distribution. Fortunately, the exact form of the wrapped-Cauchy distribution is known:

$$P_{\mathrm{WC}}(\theta) = \frac{1}{2\pi}\frac{1 - \zeta^2}{1 + \zeta^2 - 2\zeta\cos(\theta - \mu)}, \tag{3.6}$$

where $\mu = \langle\theta\rangle$ is the mean value (in our case $\mu = 0$ by symmetry), and ζ is a concentration parameter between 0 and 1, which is the wrapped analog of ϵ above. In the limit where $\zeta \to 1$ the distribution tends to a Dirac delta function $\delta_D(x)$, which is the behavior expected for identical spectra. Conversely, $\zeta = 0$ results in a uniform distribution, the behavior expected for uncorrelated spectra. A negative ζ gives distributions peaked at $\theta = \pi$ and is unphysical in this context.

3.1.3 The Probability Distribution of Phase Differences of the IGM Density

We now show that this wrapped-Cauchy form does a good job of describing the real distribution of phase differences for our simulated IGM density skewers. Note that for our simple heuristic example of randomly oriented filaments, the concentration parameter ζ only depends on the product of $kr_\perp$; whereas, in the real IGM, one expects the spectral coherence quantified by ζ to depend on the Jeans scale λ_J. Because we do not know how to directly compute the concentration parameter in terms of the Jeans scale from first principles, we opt to calculate ζ from our simulations. At any longitudinal wavenumber k, pair separation $r_\perp$, and Jeans scale λ_J, our density skewers provide a discrete sampling of the θ distribution. We use the maximum likelihood procedure from Jammalamadaka & Sengupta [2001] to calculate the best-fit value of ζ from an ensemble of θ values, as described further in Appendix B. Figure 3.2 shows the distribution of phases determined from our IGM density skewers (symbols with error bars) compared to the best-fit wrapped-Cauchy distributions (curves) for different longitudinal modes k, transverse separations $r_\perp$, and values of the Jeans scale λ_J. We see that the wrapped-Cauchy distribution typically provides a good fit to the simulation data points to within the precision indicated by the error bars. For very peaked distributions which correspond to more spectral coherence (i.e. low-k or large λ_J), there is a tendency for our wrapped-Cauchy fits to overestimate the probability of large phase differences relative to the simulated data, although our measurements of the probability are very noisy in this regime. We have visually inspected similar curves for the entire dynamic range of the relevant k, $r_\perp$ and λ_J, for which the shape of the wrapped-Cauchy distribution varies from nearly uniform ($\zeta \simeq 0$) to a very high degree of coherence ($\zeta \simeq 1$), and find similarly good agreement.

It is instructive to discuss the primary dependencies of the phase difference distribution on wavenumber k, separation $r_\perp$, and the Jeans scale λ_J illustrated in Figure 3.2. At a fixed wavenumber k, a large separation relative to the Jeans scale results in a flatter distribution of θ, which approaches uniformity for $r_\perp \gg \lambda_J$. The distribution approaches the fully coherent limit of a Dirac delta function for $r_\perp \ll \lambda_J$, and the transitions from

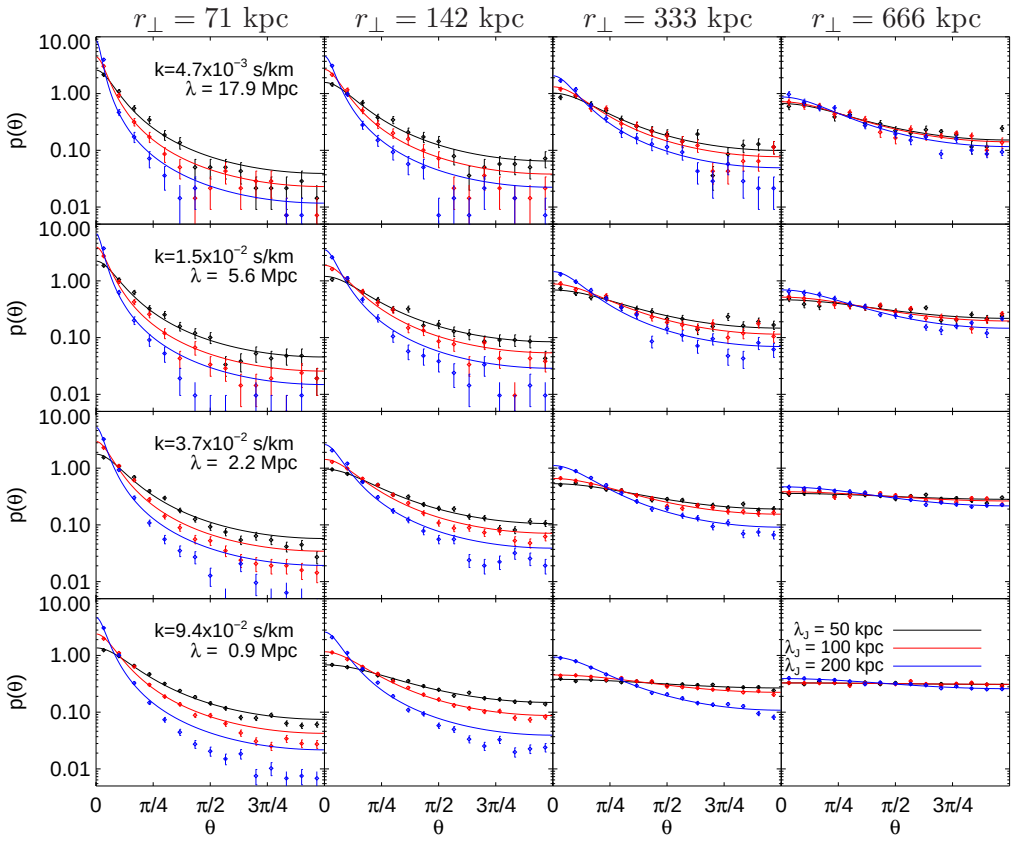

FIGURE 3.2: Phase difference probability functions of the density fields at different separations $r_\perp$, wavenumbers k and Jeans scale λ_J. Points with errorbars represent the binned phase distribution of the density field as obtained from the simulation, while the solid lines are the best-likelihood fit using a wrapped-Cauchy distribution. When the spectra are highly correlated the phases are small and the distribution is peaked around zero, whereas independent skewers result in flat probability functions. The error are estimated from the number of modes available in the simulation, assuming a Poisson distribution. By symmetry $p(\theta)$ must be even in θ, hence it is convenient to plot only the range $[0, \pi]$, summing positive and negative probabilities (clearly obtaining $p(|\theta|)$) to increase the sampling in each bin. We express the scale of each mode both giving the wavelength λ in Mpc and the wave number k (in s km^{-1}) in the transformed velocity space. The wrapped-Cauchy function traces with good approximation the phase distribution obtained from the simulation, showing less accuracy in the cases of strongly concentrated peaks, where low-probability bins are noisy. Each color is a different smoothing length: $\lambda_J = 50, 100$ and 200 kpc (respectively black, red and blue). It is important to notice that the relative distributions are different not only at scales comparable to λ_J, but also for larger modes, because the 3D power of high-k modes when projected on a 1D line contributes to all the low-k components (see the text for a detailed discussion). Secondly, it is clear that the most relevant pairs are the closest ($r_\perp \lesssim \lambda_J$), because for wide separations the coherence is too low to get useful information. These two consideration together explain why close quasar pairs are the most effective objects to measure the Jeans scale, even if they cannot be observed at high resolution.

a strongly peaked distribution to a uniform one occurs when $r_\perp$ is comparable to the Jeans scale λ_J. We see that quasar pairs with transverse separations $r_\perp \lesssim 3\lambda_J$, contain information about the Jeans scale, whereas this sensitivity vanishes for larger impact parameters. At fixed $r_\perp$, lower k-modes (i.e. larger scales) are more highly correlated (smaller θ values) as expected, because sightlines spaced closely relative to the wavelength of the mode $kr_\perp \ll 1$, probe essentially the same large scale density fluctuation. Overall, the dependencies in Figure 3.2 illustrate that there is information about the Jeans smoothing spread out over a large range of longitudinal k-modes. Somewhat surprisingly, even modes corresponding to wavelengths $\gtrsim 100$ times larger than λ_J can potentially constrain the Jean smoothing.

This sensitivity of very large-scale longitudinal k-modes to a much smaller scale cutoff λ_J in the 3D power merits further discussion. First, note that the range of wavenumbers typically probed by longitudinal power spectra of the Lyα forest lie in the range $0.005\,\mathrm{s\ km^{-1}} < k < 0.1\,\mathrm{s\ km^{-1}}$ (see Figure 2.2), corresponding to modes with wavelengths $60\,\mathrm{km\ s^{-1}} < v < 1250\,\mathrm{km\ s^{-1}}$ or $830\,\mathrm{kpc} < \lambda < 17\,\mathrm{Mpc}$. Here the low-$k$ cutoff is set by systematics related to determining the quasar continuum [see e.g. Lee, 2012], whereas the high-k cutoff is adopted to mitigate contamination of the small-scale power from metal absorption lines [McDonald et al., 2000]. In principle high-resolution (echelle) spectra FWHM$= 5\,\mathrm{km\ s^{-1}}$ probe even higher wavenumbers as large as $k \simeq 3$, however standard practice is to only consider $k \lesssim 0.1$ in model-fitting [see e.g. Zaldarriaga et al., 2001]. Thus even the highest k-modes at our disposable $k \simeq 0.1$ correspond to wavelengths $\simeq 830\,\mathrm{kpc}$ significantly larger than our expectation for the Jeans scale $\sim 100\,\mathrm{kpc}$. Furthermore, we saw in § 2.3.1 that degenerate combinations of the Jeans smoothing and the IGM temperature-density relation can produce the same small-scale cutoff in the longitudinal power. Thus both metal-line contamination and degeneracies with thermal broadening imply that while it is extremely challenging to resolve the Jeans scale spectrally, the great advantage of close quasar pairs is that they resolve the Jeans scale spatially, provided they have transverse separations $r_\perp$ comparable to λ_J. We will thus typically be working in the regime where $k/k_\perp \ll 1$, where we define $k_\perp \equiv x_0/aHr_\perp$, where $aHr_\perp$ is the transverse separation converted to a velocity and $x_0 = 2.4048$ is a constant the choice of which will become clear below.

In this regime, it is straightforward to understand why the phase differences between large-scale modes are nevertheless sensitive to the Jeans scale. Consider the quantity $\langle \cos\theta(k, r_\perp)\rangle$, which is related to the cross-power discussed in § 3.1.1. This 'moment' of the phase angle PDF can be written

$$\langle \cos\theta(k, r_\perp)\rangle = \int_{-\pi}^{\pi} P(\theta(k, r_\perp)) \cos\theta(k, r_\perp) d\theta, \tag{3.7}$$

which tends toward zero for totally uncorrelated spectra ($P(\theta) = 1/2\pi$) and towards unity for perfectly correlated, i.e. identical spectra ($P(\theta) = \delta_D(\theta)$) spectra. Following the discussion in § 3.1.1, we can write

$$\begin{aligned} \pi(k, r_\perp) = \langle \rho_1(k)\rho_2(k) \cos\theta(k, r_\perp)\rangle &\approx \\ \langle \rho_1(k)\rho_2(k)\rangle\langle \cos\theta(k, r_\perp)\rangle &\approx P(k)\langle \cos\theta(k, r_\perp)\rangle, \end{aligned} \tag{3.8}$$

where the first approximation is a consequence of the approximate Gaussianity of the density fluctuations, and the second from the fact that $\langle \rho_1\rho_2\rangle \approx P(k)$ for $k/k_\perp \ll 1$, as demonstrated by the dashed curves in the right panel of Fig 2.2. Thus we arrive at

$$\langle \cos\theta(k, r_\perp)\rangle \approx \frac{\pi(k, r_\perp)}{P(k)} = \frac{\int_k^\infty dqqJ_0(r_\perp\sqrt{q^2 - k^2})P_{3D}(q)}{\int_k^\infty dqqP_{3D}(q)}, \tag{3.9}$$

where J_0 is the cylindrical Bessel function of order zero. The numerator and denominator of the last equality in eqn. (3.9) follow from the definitions of the longitudinal and cross power for an isotropic 3D power spectrum [see e.g. Hui et al., 1999, Lumsden et al., 1989, Peacock, 1999, Viel et al., 2002]. The denominator is the familiar expression for the 1D power expressed as a projection of the 3D power. Note that 1D modes with wavenumber k receive contributions from all 3D modes with wavevectors $\geq k$, which results simply from the geometry of observing a 3D field along a 1D skewer. A long-wavelength (low-k) 1D longitudinal mode can be produced by a short-wavelength (high-k) 3D mode directed nearly perpendicular to the line of sight [see e.g. Peacock, 1999]. The numerator of eqn. (3.9) is similarly a projection over all high-k 3D modes, but because of the non-zero separation of the skewers the 3D power spectrum is now modulated by the cylindrical Bessel function $J_0(x)$. Because $J_0(x)$ is highly oscillatory, the primary contribution to this projection integral will come from arguments in the range $0 < x < x_0$. Here $x_0 = 2.4048$ is the first zero of $J_0(x)$, which motivates our earlier definition of $k_\perp \equiv x_0/aHr_\perp$. For larger arguments x, the decay of $J_0(x)$ and its rapid oscillations will result in cancellation and negligible contributions. Thus for $k/k_\perp \ll 1$, we can finally write

$$\langle \cos\theta(k, r_\perp)\rangle \approx \frac{\int_k^{k_\perp} dqqJ_0(r_\perp\sqrt{q^2 - k^2})P_{3D}(q)}{\int_k^\infty dqqP_{3D}(q)}. \tag{3.10}$$

This equation states that the average value of the phase difference between homologous k modes is determined by the ratio of the 3D power integrated against a 'notch filter' which transmits the range $[k, k_\perp]$, relative to the total integrated 3D power over the full range $[k, \infty]$. Hence phase angles between modes with wavelengths $\gtrsim 100$ times larger than λ_J, are nevertheless sensitive to the amount of 3D power down to scales as small as the transverse separation $r_\perp$. This results simply from the geometry of observing a 3D field

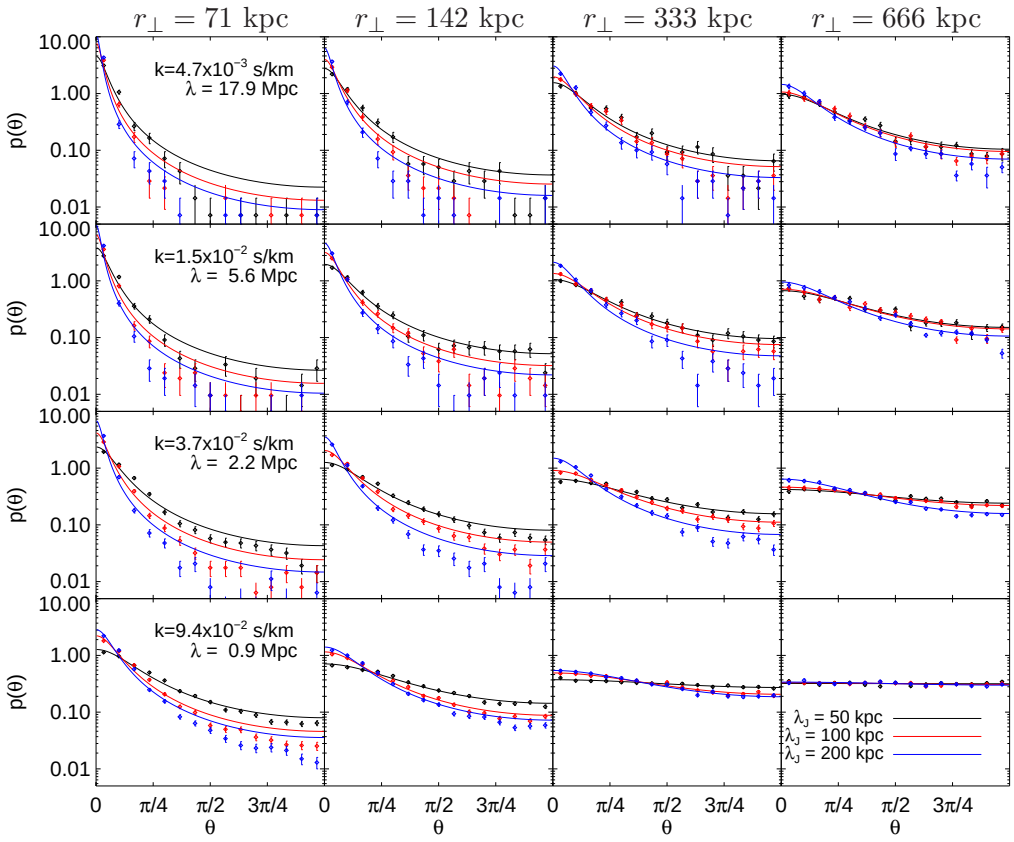

FIGURE 3.3: Same plot of figure 3.2 but for the Lyα transmitted flux field instead of density. We vary the Jeans scale λ_J, keeping fixed the equation-of-state parameters, $T_0 = 10000$ K and $\gamma = 1.6$. The properties of the distributions are analogous to the previous plot, they follow with good approximation a wrapped-Cauchy profile and they exhibit the same trends with $r_\perp, k$ and λ_J. Overall, the flux shows an higher degree of coherence and a slightly smaller sensitivity to λ_J.

along 1D skewers, because the power in longitudinal mode k is actually dominated by the superposition of 3D power from much smaller scales $\gg k$. Provided that quasar pair separations resolve the Jeans scale $r_\perp \sim \lambda_J$, even large scale modes with $k \ll k_\perp \sim 1/\lambda_J$ are sensitive to the shape of the 3D power on small-scales, which explains the sensitivity of low-k modes to the Jeans scale in Figure 3.2.

Finally, the form of eqn. (3.10) combined with eqn. (3.7) explains the basic qualitative trends in Figure 3.2. For large $r_\perp$ (small $k_\perp$) the projection integral in the numerator decreases, $\langle \cos\theta(k, r_\perp)\rangle$ approaches zero, indicating that $P(\theta(k, r_\perp))$ approaches uniformity. Similarly, as $r_\perp \to \lambda_J$, $\langle \cos\theta(k, r_\perp)\rangle$ grows indicating that $P(\theta(k, r_\perp))$ is peaked toward small phase angles, and in the limit $r_\perp \ll \lambda_J$ $\langle \cos\theta(k, r_\perp)\rangle \to 1$ and $P(\theta(k, r_\perp))$ approaches a Dirac delta function. At fixed $r_\perp$, lower k modes will result in more common pathlength in the projection integrals in the numerator and denominator of eqn. (3.10), thus $\langle \cos\theta(k, r_\perp)\rangle$ is larger, $P(\theta(k, r_\perp))$ is more peaked, and the phase angles are more highly correlated.

To summarize, following a simple heuristic argument, we derived a analytical form for the phase angle distribution in § 3.1.2, which is parametrized by a single number, the concentration ζ. We verified that this simple parametrization provides a good fit to the distribution of phase differences in our simulated skewers, and explored the dependence of this distribution on transverse separation $r_\perp$, wavenumber k, and the Jeans scale λ_J. Phase differences between large-scale modes with small wavenumbers $k \ll 1/\lambda_J$, are sensitive to the Jeans scale, because geometry dictates that low-k cross-power across correlated 1D skewers is actually dominated by high-k 3D modes up to a scale set by the pair separation $k_\perp \sim 1/r_\perp$.

3.1.4 The Probability Distribution of Phase Differences of the Flux

Having established that the wrapped-Cauchy distribution provides a good description of the phase difference of IGM density skewers, we now apply it to the Lyα forest flux. Figure 3.3 shows the PDF of phase differences for the exact same transverse separations $r_\perp$, wavenumbers k, and Jeans smoothings λ_J that were shown in Figure 3.2. The other thermal parameters T_0 and γ have been set to $(T_0, \gamma) = (10,000\,\mathrm{K}, 1.6)$. Overall, the behavior of the phase angle PDF for the flux is extremely similar to that of the density, exhibiting the same basic trends. Namely, the flux PDF also transitions from a strongly peaked distribution ($r_\perp \lesssim \lambda_J$) to a flat one ($r_\perp \gg \lambda_J$) at around $r_\perp \simeq \lambda_J$. Lower k-modes tend to be more highly correlated, and low-k modes corresponding to wavelengths $\gtrsim 100\lambda_J$ are nevertheless very sensitive to the Jeans scale, in exact analogy with the density field. Note that because the 3D power spectrum of the flux field is now anisotropic, the assumptions leading to the derivation of eqn. 3.10 in the previous section breaks down for the flux. Nevertheless, the explanation for the sensitivity of low-k modes to the Jeans scale is likely the same, namely the low-k power across correlated skewers is actually dominated by projected high-k 3D power up to a scale $k_\perp \sim 1/r_\perp$, which is set by the pair separation.

The primary difference between the phase angle PDF of flux versus the density appears to be that the flux PDF is overall slightly less sensitive to the Jeans scale. In general, we do not expect the two distributions to be exactly the same for several reasons. First, the flux represents a highly nonlinear transformation of the density: according to the FGPA formula $\delta F \sim \exp\left[-(1+\delta)^\beta\right]$ where $\beta = 2 - 0.7(\gamma - 1)$. Second, the flux is observed in redshift space, and the peculiar velocities which determine the mapping from real to redshift space, can further alter the flux relative to the density. Finally, the flux field is sensitive to other thermal parameters T_0 and γ, both through the nonlinear FGPA transformation, and because of thermal broadening. In what follows, we investigate

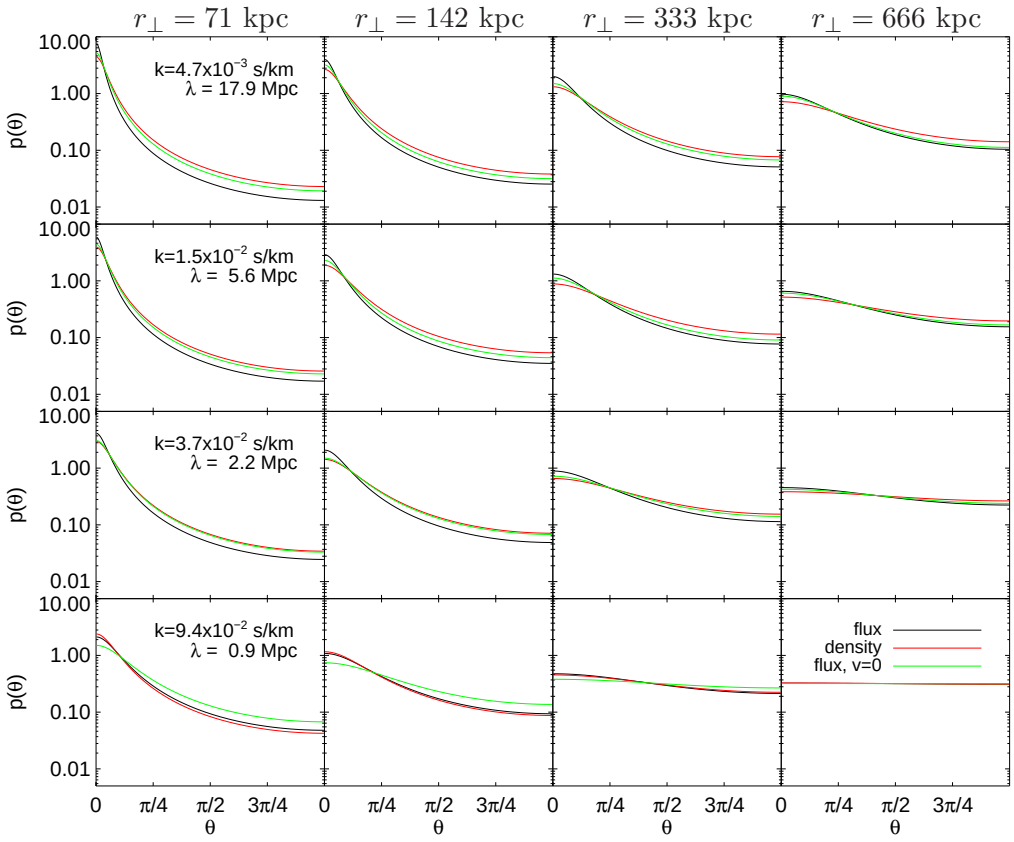

FIGURE 3.4: Phase difference probability density functions for different separations $r_\perp$ and wavenumbers k. All models have the same Jeans scale $\lambda_J = 140$ kpc. For clarity we plot only the best-fit wrapped-Cauchy function without simulated points with errorbars. The black and the red lines are the phase angle PDFs for the transmitted flux of the Lyαforest and the IGM density field, respectively. The green line represents the case of the Lyαforest flux where peculiar velocities are set to zero. By comparing the green and the black lines we see that in peculiar motions always increase the coherence between the two sightlines, which partly explains the differences between the flux and density distributions, since the latter is calculated in real space. The flux and density further differ because of the non-linear FGPA transformation, which has a stronger effect on smaller scale modes.

each of these effects in turn, and discuss how each alters the phase angle PDF and its sensitivity to the Jeans scale.

In Figure 3.4 we show the flux PDF (black) alongside the density PDF (red) for various modes and separations, again with the thermal model fixed to $(T_0, \gamma, \lambda_J) = (20,000\ \text{K}, 1.0, 140)$ kpc. To isolate the impact of peculiar velocities, we also compute the phase angle PDF of the real-space flux, i.e. without peculiar velocities (green). Specifically, we disable peculiar velocities by computing the flux from eqn. (2.6) with $v_{p,\parallel}$ set to zero. Overall, the PDFs of the real-space flux and density (also real-space) are quite similar. For low wavenumbers, the real-space flux skewers are always slightly more coherent than the density ($P(\theta)$ more peaked) for all separations. However, at the highest k, the situation

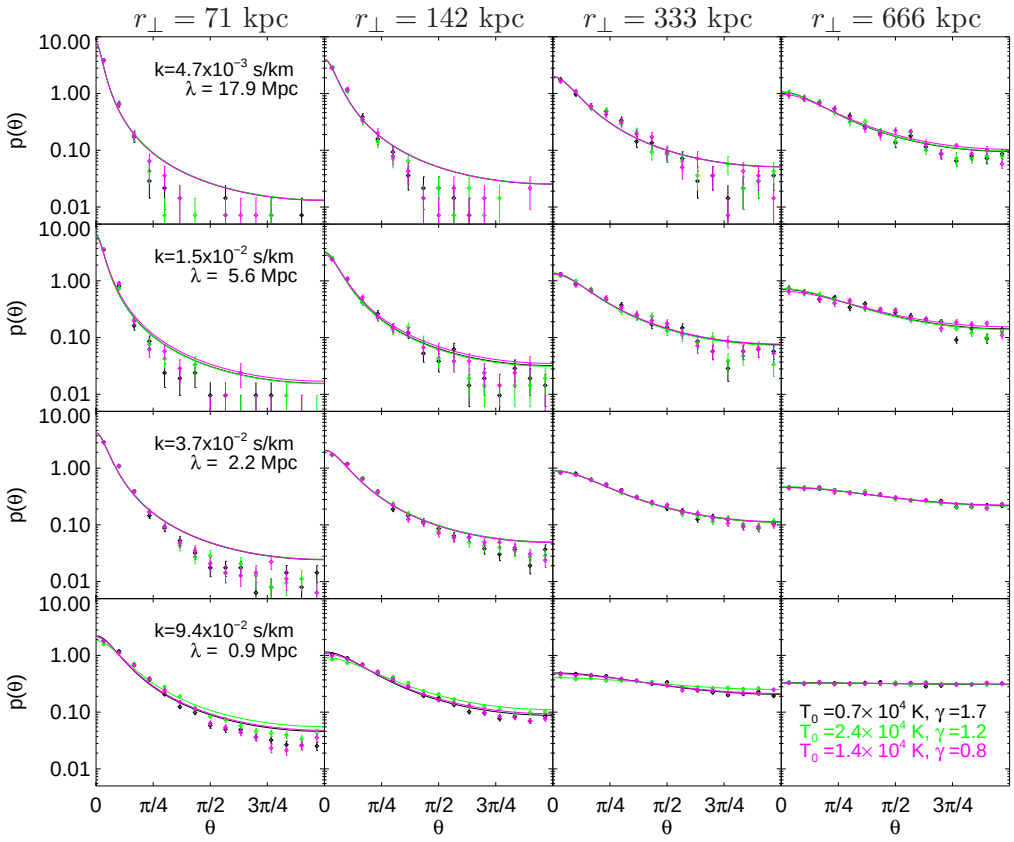

FIGURE 3.5: Phase difference probability density functions for different separations $r_\perp$, wavenumbers k and equation-of-state parameters $T_0 - \gamma$. Points with errorbars (estimated Poisson error) are the results of our simulations, while the coloured lines are the best-likelihood fit using a wrapped-Cauchy distribution. All models have the same Jeans scale $\lambda_J = 140$ kpc. This plot shows the most remarkable property of phases: they do not exhibit any relevant sensitivity to the equation of state, so they robustly constrain the spatial coherence given by pressure support.

is reversed with the density being more coherent than the real-space flux. A detailed explanation of the relationship between the phase angle PDF of the real-space flux and the density fields requires a better understanding of the effect of the non-linear FGPA transformation on the 2-point function of the flux, which is beyond the scope of the present work. Here we only argue that the 3D power spectrum of the real-space flux has in general a different shape than that of the density, and using our intuition from eqn. (3.10), this will result in a different shape for the distribution of phase angles. The net effect of peculiar velocities on the redshift-space flux PDF is to increase the amount of coherence between the two sightlines ($P(\theta)$ more peaked) relative to the real-space flux. This likely arises because the peculiar velocity field is dominated by large-scale power, which makes the 3D power of the flux steeper as a function of k. Again based on our intuition from eqn. (3.9), a steeper power spectrum will tend to increase the coherence ($\langle\cos(\theta(k, r_\perp))\rangle$ closer to unity), because the projection integrals in the numerator and denominator of eqn. (3.9) will both have larger relative contributions from

the interval $[k, k_\perp]$. Note that the relative change in the flux PDF due to peculiar velocities is comparable to the differences between the real-space flux and the density. At the highest k-values where the real-space flux is less coherent than the density (lowest panel of Figure 3.4), peculiar velocities conspire to make the redshift-space flux PDF very close to the density PDF.

Finally, we consider the impact of the other thermal parameters T_0 and γ on the distribution of phases in Figure 3.5. There we show the PDF of the phase angles for the flux for a fixed Jeans scale $\lambda_J = 140\,\mathrm{kpc}$, and three different thermal models. Varying T_0 and γ over the full expected range of these parameters has very little impact on the shape of the phase angle PDF, whereas we see in Figure 3.3 that varying the Jeans scale has a much more dramatic effect. The physical explanations for the insensitivity to T_0 and γ are straightforward. The thermal parameters T_0 and γ can influence the phase angle PDF in two ways. First, the FGPA depends weakly on temperature $T^{-0.7}$ through the recombination coefficient. As a result the non-linear transformation between density and flux depends weakly on γ $\delta F \sim \exp\left[-(1+\delta)^\beta\right]$ where $\beta = 2 - 0.7(\gamma - 1)$. We speculate that the tiny differences between the thermal models in Figure 3.5 are primarily driven by this effect, because we saw already in Figure 3.5 that the non-linear transformation can give rise to large differences between the density and flux PDFs. This small variation of the PDF with γ then suggests that it is actually the exponentiation which dominates the differences between the flux and density PDFs in Figure 3.5, with the weaker γ dependent transformation $(1+\delta)^{2-0.7(\gamma-1)}$ playing only a minor role, which is perhaps not surprising. Note that there is also a $T_0^{-0.7}$ dependence in the coefficient of the FGPA optical depth, but as we require all models to have the same mean flux $\langle\exp(-\tau)\rangle$, this dependence is compensated by the freedom to vary the metagalactic photoionization rate Γ. Second, both T_0 and γ determine the temperature of gas at densities probed by the Lyα forest, which changes the amount of thermal broadening. The insensitivity to thermal broadening is also rather easy to understand. Thermal broadening is effectively a convolution of the flux field with a Gaussian smoothing kernel. In k-space this is simply a multiplication of the Fourier transform of the flux $\delta\tilde{F}(k)$ with the Fourier transform of the kernel. Because all symmetric kernels will have a vanishing imaginary part[1], the convolution can only modify the moduli of the flux *but the phases are invariant.* Thus the phase differences between neighboring flux skewers are also invariant to smoothing, which explains the insensitivity of the flux phase angle PDF to thermal broadening, and hence the parameters T_0 and γ.

The results of this section constitute the cornerstones of our method for measuring the Jeans scale. We found that the phase angle PDF of the flux has a shape very similar to

[1]The imaginary part of the Fourier transform of the symmetric function $W(|x|)$ is $\Im[W(k)] = \int W(|x|)\sin(kx)dx$ which is always odd and will integrate to zero.

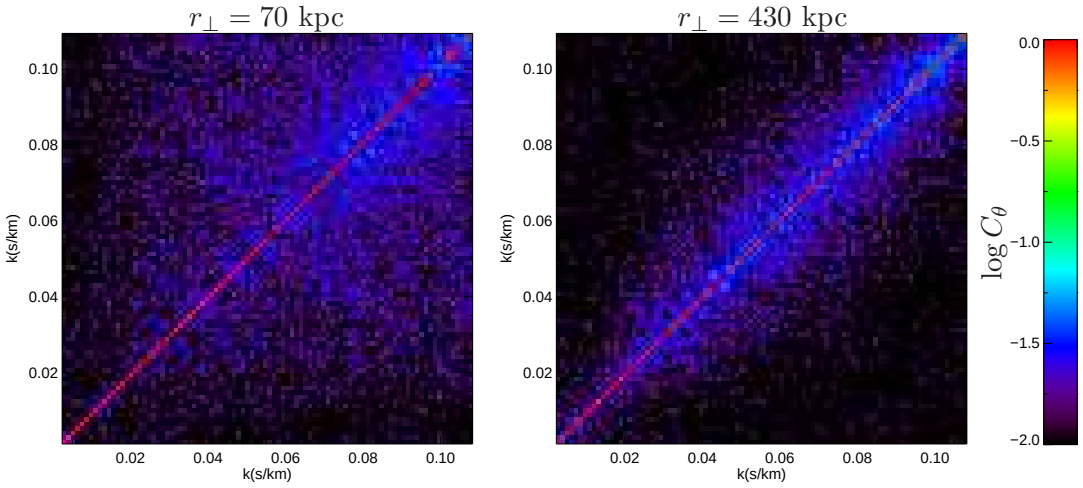

FIGURE 3.6: Logarithm of the phase $k - k$ correlation for separations $r_\perp = 70$ kpc (left) and $r_\perp = 430$ kpc (right). This matrices are calculated for a model with $\lambda_J = 143$ kpc, $T_0 = 20000$ K and $\gamma = 1$. Phases are more correlated when the impact parameter is smaller than the jeans scale and at high k where nonlinear growth of perturbations couples different modes. Even in this cases we rarely find correlations higher than $\approx 3\%$, for which reason we will work in the diagonal approximation. This approximation may break out if the measured Jeans scale will be significantly larger than expected.

that of the density, and that both are well described by the single parameter wrapped-Cauchy distribution. Information about the 3D smoothing of the density field λ_J, is encoded in the phase angle PDF of the flux, but it is essentially independent of the other thermal parameters governing the IGM. This results because 1) the non-linear FGPA transformation is only weakly dependent on temperature 2) phase angles are invariant under symmetric convolutions. The implication is that close quasar pair spectra can be used to pinpoint the Jeans scale without suffering from any significant degeneracies with T_0 and γ. Indeed, in the next section we introduce a Bayesian formalism for estimating the Jeans scale, and our MCMC analysis in § 3.2 will assess the accuracy with which the thermal parameters can be measured, and explicitly demonstrate the near independence of constraints on λ_J from T_0 and γ.

3.1.5 The Covariance of the Phase Differences

In the previous section, we showed that the PDF of phase differences between homologous longitudinal modes of the flux field are well described by the wrapped-Cauchy distribution (see eqn. 3.6). However, the one-point function alone is insufficient for characterizing the statistical properties of the stochastic field $\theta(k, r_\perp)$, because in principle values of θ closely separate in either wavenumber k or real-space could be correlated. Understanding the size of these two-point correlations is of utmost importance. Any given quasar pair spectrum provides us with a realization of $\theta(k, r_\perp)$, and we have seen

that the distribution of these values depends sensitively on the Jeans scale λ_J. In order to devise an estimator for the thermal parameters in terms of the phase differences, we have to understand the degree to which the $\theta(k, r_\perp)$ are independent.

It is easy to rule out the possibility of spatial correlations among the θ values deduced from distinct quasar pairs. Because quasar pairs are extremely rare on the sky, the individual quasar pairs in any observed sample will typically be $\sim$ Gpc away from each other, and hence different pairs will never probe correlated small-scale density fluctuations. However, the situation is much less obvious when it comes to correlations between θ values for different k-modes of the same quasar pair. In particular, nonlinear structure formation evolution will result in mode-mode coupling, which can induce correlations between mode amplitudes and phases [e.g. Chiang et al., 2002, Coles, 2009, Watts et al., 2003]. We are thus motivated to use our simulated skewers to directly quantity the size of the correlations between phase differences of distinct longitudinal k-modes.

We calculate the correlation coefficient matrix of θ between modes k and k' defined as

$$C_\theta(k, k'; r_\perp) = \frac{\langle \theta(k, r_\perp)\theta(k', r_\perp)\rangle}{\sqrt{\langle \theta^2(k, r_\perp)\rangle \langle \theta^2(k', r_\perp)\rangle}}. \tag{3.11}$$

Our standard setup of 330 pairs at each discrete separation $r_\perp$ results in a very noisy estimate of $C_\theta(k, k'; r_\perp)$, so we proceed by defining a new set of 80,000 skewers at two distinct discrete transverse separations of $r_\perp = 70$ kpc and $r_\perp = 430$ kpc for a single thermal model with $(T_0, \gamma, \lambda_J) = (20,000\,\mathrm{K}, 1, 143\,\mathrm{kpc})$.

Figure 3.6 displays the correlation coefficient matrix for the two separations $r_\perp$ that we simulated. We find that the off-diagonal correlations between k-modes are highest at high k values and for smaller impact parameters. This is the expected behavior, since higher longitudinal k-modes will have a larger relative contributions from higher-k 3D modes, which will be more non-linear and have larger mode-mode correlations. Likewise, as per the discussion in § 3.1.3, phase differences at smaller pair separations $r_\perp$ are sensitive to higher k 3D power $\sim k_\perp$, and should similarly exhibit larger correlations between modes. Note however that over the range of longitudinal k values which we will use to constrain the Jeans scale $0.005 < k < 0.1$, the size of the off-diagonal elements are always very small, of the order of $\sim 1 - 3\%$.

The small values of the off-diagonal elements indicates that the mode-mode coupling resulting from non-linear evolution does not result in significant correlations between the phase angles of longitudinal modes. This could result from the fact that the intrinsic phase correlations of the 3D modes is small, and it is also possible that the projection of power inherent to observing along 1D skewers (see § 3.1.3) dilutes these intrinsic phase correlations, because a given longitudinal mode is actually the average over a large

range of 3D modes. From a practical perspective, the negligible off-diagonal elements in Figure 3.6 are key, because they allow us to consider each phase difference $\theta(k, r_\perp)$ as an *independent* random draw from the probability distributions we explored in § 3.1.4, which as we show in the next section, dramatically simplifies the estimator that we will use to determine the Jeans scale.

3.1.6 A Likelihood Estimator for the Jeans Scale

The results from the previous sections suggest a simple method for determining Jeans scale. Namely, given any quasar pair, the phase angle difference for a given k-mode represents a draw from the underlying phase angle PDF determined by the thermal properties of the IGM (as well as other parameters governing e.g. cosmology and the dark matter which we assume to be fixed). In § 3.1.4 we showed that the phase angle PDF is well described by the wrapped-Cauchy distribution and in § 3.1.5 we argued that correlations between phase angle differences $\theta(k, r_\perp)$, in both k-space and real-space can be neglected. Thus for a hypothetical dataset $\theta(k, r_\perp)$ measured from a sample of quasar pairs, we can write that the likelihood of the thermal model $M = \{T_0, \gamma, \lambda_J\}$ given the data is

$$\mathscr{L}(\{\theta\}|M) = \prod_{i,j} P_{\rm WC}(\theta(k_i, r_j)|\zeta(k, r_\perp|M)). \tag{3.12}$$

This states that the likelihood of the data is the product of the phase angle PDF evaluated at the measured phase differences for all k-modes and over all quasar pair separations $r_\perp$. Note that the simplicity of this estimator is a direct consequence of the fact that there are negligible θ correlations between different k-modes and pair separations. All dependence on (T_0, γ, λ_J) is encoded in the single parameter ζ, which is the concentration of the wrapped-Cauchy distribution (eqn. 3.6).

We can then apply Bayes' theorem to make inferences about any thermal parameter, for example for λ_J

$$P(\lambda_J|\{\theta\}) = \frac{\mathscr{L}(\{\theta\}|\lambda_J)p(\lambda_J)}{P(\{\theta\})} \tag{3.13}$$

where $p(\lambda_J)$ is our prior on the Jeans scale and the denominator acts as a renormalization factor which is implicitly calculated by a Monte Carlo simulation over the parameter space. The same procedure can be used to evaluate the probability distribution of the other parameters. Throughout this work, we assume flat priors on all thermal parameters, over the full domain of physically plausible parameter values.

In § 3.2 we will use MCMC techniques to numerically explore the likelihood in eqn. (3.13) and deduce the posterior distributions of the thermal parameters. In order to do this, we need to be able to evaluate the function $\zeta(k, r_\perp|T_0, \gamma, \lambda_J)$ at any location in thermal

parameter space. This is a non-trivial computational issue, because we do not have a closed form analytical expression for ζ which can be evaluated quickly, and thus have to resort to our cosmological simulations of the IGM to numerically determine it for each model, as described in Appendix B. In practice, computational constraints limit the size of our thermal parameter grid to only 500 thermal models, and we thus evaluate ζ at only these 500 fixed locations. The fast procedure described in the previous chapter (the *emulator*) allows us to interpolate ζ from these 500 locations in our finite thermal parameter grid, onto any value in thermal parameter space (T_0, γ, λ_J).

To summarize, our method for measuring the Jeans scale of the IGM involves the following steps:

- Calculate the phase differences $\theta(k, r_\perp)$ for each k-mode of an observed sample of quasar pairs with separations $r_\perp$.
- Generate Lyα forest quasar pair spectra for a grid of thermal models in the parameter space (T_0, γ, λ_J), using our IGM simulation framework. For each model, numerically determine the concentration parameter $\zeta(k, r_\perp | T_0, \gamma, \lambda_J)$ at each wavenumber k and separation $r_\perp$, from the distribution of phase differences $\theta(k, r_\perp)$.
- Emulate the function $\zeta(k, r_\perp | T_0, \gamma, \lambda_J)$, enabling fast interpolation of ζ from the fixed values in the thermal parameter grid to any location in thermal parameter space.
- Calculate the posterior distribution in eqn. (3.13) for λ_J, by exploring the likelihood function in eqn. (3.12) with an MCMC algorithm.

3.2 How Well Can We Measure the Jeans Scale?

Our goal in this section is to determine the precision with which close quasar pair spectra can be used to measure the Jeans scale. To this end, we construct a mock quasar pair dataset from our IGM simulations and apply our new phase angle PDF likelihood formalism to it. A key question is how well constraints from our new phase angle technique compare to those obtainable from alternative measures, such as the cross-power spectrum, applied to the same pair sample, or from the longitudinal power spectrum, measured from samples of individual quasars. In what follows, we first present the likelihood used to determine thermal parameter constraints for these two additional statistics. Then we describe the specific assumptions made for the mock data. Next we quantify the resulting precision on the Jeans scale, explore degeneracies with other thermal parameters, and compare to constraints from these two alternative statistics. We explore the

impact of finite signal-to-noise ratio and spectral resolution on our measurement accuracy, and discuss possible sources of systematic error. Finally, we explicitly demonstrate that our likelihood estimator provides unbiased determinations of the Jeans scale.

3.2.1 The Likelihood for $P(k)$ and $\pi(k, r_\perp)$

For the longitudinal power $P(k)$, we assume that the distribution of differences, between the measured band powers of a k-bin and the true value, is a multi-variate Gaussian [see e.g. McDonald et al., 2006], which leads to the standard likelihood for the power-spectrum

$$\begin{aligned} \mathscr{L}(P_d|M) &= (2\pi)^{-N/2} \det(\Sigma)^{-1/2} \\ &\quad \exp\left[-\frac{1}{2}(P_d - P_M)^T \Sigma^{-1} (P_d - P_M)\right], \end{aligned} \tag{3.14}$$

where P_d is a vector of N observed 1D band powers, P_M is a vector of power spectrum predictions for a given thermal model $M = (T_0, \gamma, \lambda_J)$, and

$$\Sigma(k, k') = \langle [P(k) - P_M(k)][P(k') - P_M(k')] \rangle, \tag{3.15}$$

is the full covariance matrix of the power spectrum measurement. As we describe in the next subsection, we will choose a subset of the skewers from a fiducial thermal model to represent the 'data' in this expression, which are then compared directly to thermal models (T_0, γ, λ_J), where the same emulator technique described in § 2.2 is used to interpolate $P_M(k|T_0, \gamma, \lambda_J)$ to parameter locations in the thermal space. To determine the covariance of this mock data $\Sigma(k, k')$, we use the full ensemble of $2 \times 10,000$ 1D skewers for the fiducial thermal model, directly evaluate the covariance matrix, and then rescale it to the size of our mock dataset by multiplying by the ratio of the diagonal terms $\sigma^2_{\rm dataset}/\sigma^2_{\rm full}$. This procedure of evaluating the covariance implicitly assumes that the only source of noise in the measurement is sample variance, or that the instrument noise is negligible. For the high-resolution and high signal-to-noise ratio spectra used to measure the longitudinal power spectrum cutoff [Croft et al., 2002, McDonald et al., 2000], this is a reasonable assumption. For reference, the relative magnitude of off-diagonal terms of the covariance, $\Sigma(k, k')/\sqrt{\Sigma(k, k)\Sigma(k', k')}$, are at most $20 - 30\%$ with the largest values attained at the highest k.

For the cross-power spectrum $\pi(k, r_\perp)$, we follow the same procedure. Namely, a mock dataset is constructed for the fiducial thermal model by taking a subset of the full ensemble of quasar pair spectra. We again assume that the band power errors are distributed according to a multi-variate Gaussian, but because we must now account for

the variation with separation $r_\perp$, the corresponding likelihood is

$$\mathscr{L}(\pi|M) = \prod_i \mathscr{L}(\pi_d(k, r_{\perp,i})|M), \tag{3.16}$$

where $\mathscr{L}(\pi_d(k, r_{\perp,i})|M)$ has the same form as the longitudinal power in eqn. (3.15). In exact analogy with the longitudinal power, we compute the full covariance matrix $\Sigma(k, k'|r_\perp)$ of the cross-power using our full ensemble of simulated pair spectra for our fiducial model, but now at each value of $r_\perp$.

3.2.2 Mock Datasets

To determine the accuracy with which we can measure the Jeans scale and study the degeneracies with other thermal parameters, we construct a dataset with a realistic size and impact parameter distribution, and use an MCMC simulation to explore the phase angle likelihood in eqn. (3.12). We compare these constraints to those obtained from the cross-power spectrum for the same mock pair dataset, by similarly using an MCMC to explore the cross-power likelihood in eqn. (3.16). We also compare to parameter constraints obtainable from the longitudinal power alone, by exploring the likelihood in eqn. (3.15), for which we must also construct a mock dataset for longitudinal power measurements.

For the mock quasar pair sample, we assume 20 quasar pair spectra at $z = 3$, with fully overlapping absorption pathlength between Lyα and Lyβ. Any real quasar pair sample will be composed of both binary quasars with full overlap and projected quasar pairs with partial overlap, so in reality 20 represents the total effective pair sample, whereas the actual number of quasar pairs required could be larger. The distribution of transverse separations for these pairs is taken to be uniform in the range $24 < r_\perp < 714$ kpc. Specifically, we require 200 pairs of skewers in order to build up the necessary path length for 20 full Lyα forests, and these are randomly selected from our 10,000 IGM pair skewers which have 30 discrete separations. We draw these pairs from a simulation with a fiducial thermal model $(T_0, \gamma, \lambda_J) = (12,000\,\mathrm{K}, 1.0, 110,\ \mathrm{kpc})$, which lies in the middle of our thermal parameter grid. Note that follow-up observations of quasar pair candidates has resulted in samples of > 400 quasar pairs in the range $1.6 < z \lesssim 4.3$ with $r_\perp < 700\,\mathrm{kpc}$, and for those with $> 50\%$ overlap, the total effective number of fully overlapping pairs is $\simeq 300$ [Hennawi, 2004, Hennawi et al., 2009, 2006b, Myers et al., 2008]. Many of these sightlines already have the high quality Lyα forest spectra required for a Jeans scale measurement [e.g. Hennawi & Prochaska, 2007, 2008, Hennawi et al., 2006a, Prochaska & Hennawi, 2009, Prochaska et al., 2012], hence the mock dataset

we have assumed already exists, and can be easily enlarged given the number of close quasar pairs known.

Longitudinal power spectrum measurements which probe the small-scale cutoff of the power have been performed on high-resolution ($R \simeq 30,000 - 50,000$; FWHM=$6 - 10\,\mathrm{km\ s^{-1}}$) spectra of the brightest quasars. Typically, the range of wavenumbers used for model fitting is $0.005\,\mathrm{s\ km^{-1}} < k < 0.1\,\mathrm{s\ km^{-1}}$ (see Figure 2.2), where the low-k cutoff is chosen to avoid systematics related to quasar continuum fitting [Lee, 2012], and the high-k cutoff is adopted to mitigate contamination from metal absorption lines [Croft et al., 2002, Kim et al., 2004, McDonald et al., 2000]. Because quasar pairs are very rare, one must push to faint magnitudes to find them in sufficient numbers. In contrast with the much brighter quasars used to measure the small-scale longitudinal power [Croft et al., 2002, Kim et al., 2004, McDonald et al., 2000], quasar pairs are typically too faint to be observable at echelle resolution (FWHM=$6 - 10\,\mathrm{km\ s^{-1}}$) on 8m class telescopes. However, quasar pairs can be observed with higher efficiency echellette spectrometers, which deliver $R \simeq 10,000$ or FWHM$= 30\,\mathrm{km\ s^{-1}}$. The cutoff in the power spectrum induced by this lower resolution is $k_{\mathrm{res}} = 1/\sigma_{\mathrm{res}} = 2.358/\mathrm{FWHM} = 0.08\,\mathrm{s\ km^{-1}}$, which is very close to the upper limit $k < 0.1\,\mathrm{s\ km^{-1}}$ set by metal-line contamination. For these reasons, we will consider only modes in the range $0.005\,\mathrm{s\ km^{-1}} < k < 0.1\,\mathrm{s\ km^{-1}}$ in the likelihood in eqn. (3.12). We initially consider perfect data, ignoring the effect of finite signal-to-noise rate and resolution. Then in § 3.2.4, we will explore how noise and limited spectral resolution influence our conclusions.

For the mock sample used to study the longitudinal power, we assume perfect data, which is reasonable considering that such analyses are typically performed on spectra with signal-to-noise ratio S/N ~ 30 and resolution FWHM$= 6\,\mathrm{km\ s^{-1}}$ [Croft et al., 2002, Kim et al., 2004, McDonald et al., 2000] such that the Lyα forest, and in particular modes with $k < 0.1$, are fully resolved. For the size of this sample, we again assume 20 individual spectra at $z = 3$ with full coverage of the Lyα forest, which is about twice the size employed in recently published analyses [Croft et al., 2002, Kim et al., 2004, McDonald et al., 2000]. However, the number of existing archival high-resolution quasar spectra at $z = 3$ easily exceeds this number, so samples of this size are also well within reach. Also, adopting a sample for the longitudinal power with the same Lyα forest path length as the quasar pair sample, facilitates a straightforward comparison of the two sets of parameter constraints.

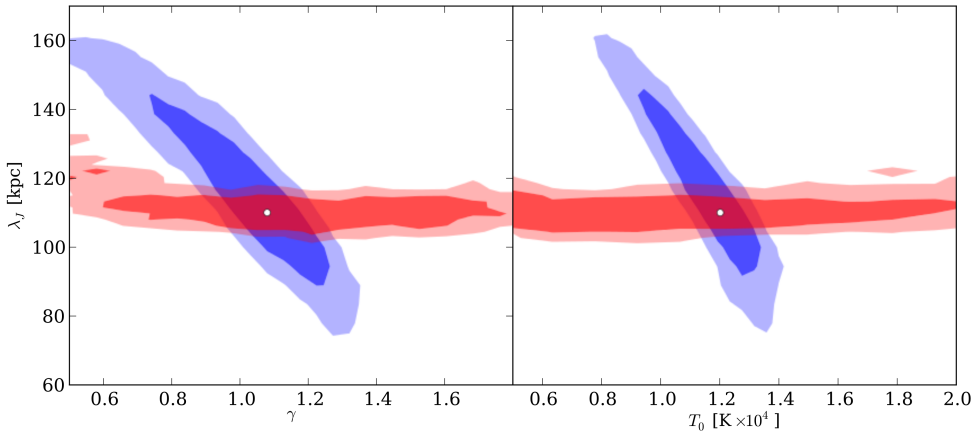

FIGURE 3.7: Constraints on the $\gamma - \lambda_J$ and $T_0 - \lambda_J$ planes. The contours show the estimated 65% and 96% confidence levels obtained with the longitudinal power (blue) and the phase difference (red). The white dot marks the fiducial model in the parameter space. The degeneracy affecting the 1D power already shown in figure 2.2 can now be seen clearly in the parameter space through the inclination of the black contours. Conversely, the fact that constraints given by the phase difference statistic are horizontal guarantees that this degeneracy is broken and that the measurement of the Jeans scale is not biased by the uncertainties on the equation of state.

3.2.3 The Precision of the λ_J Measurement

Given our mock dataset and the expression for the phase angle likelihood in eqn. (3.12), and armed with our IGM emulator, which enables us to quickly evaluate this likelihood everywhere inside our thermal parameter space, we are now ready to explore this likelihood with an MCMC simulation to determine the precision with which we can measure the Jeans scale and explore degeneracies with other thermal parameters.

We employ the publicly available MCMC package described in Foreman-Mackey et al. [2012], which is particularly well adapted to explore parameter degeneracy directions. The result of our MCMC simulation is the full posterior distribution in the 3-dimensional $T_0 - \gamma - \lambda_J$ space for each likelihood that we consider. It is important to point out that, in general, these posterior distributions will not be exactly centered on the true fiducial thermal model $(T_0, \gamma, \lambda_J) = 12,000\,\mathrm{K}, 1, 110,$ kpc. Indeed, the expectation is that the mean or mode of the posterior distribution for a given parameter will scatter around the true fiducial value at a level comparable to the width of this distribution. Nevertheless, the posterior distribution should provide an accurate assessment of the precision with which parameters can be measured and the degeneracy directions in the parameter space. In § 3.2.6 we will demonstrate that our phase angle PDF likelihood procedure is indeed an unbiased estimator of the Jeans scale, by applying this method to a large ensemble of mock datasets, and showing that on average, we recover the input fiducial Jeans scale.

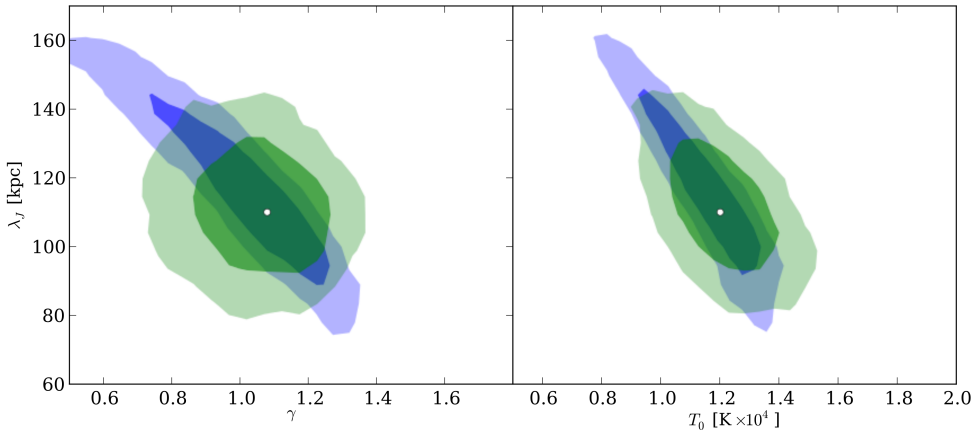

FIGURE 3.8: Constraints on the $\gamma - \lambda_J$ and $T_0 - \lambda_J$ planes. The contours show the estimated 65% and 96% confidence levels obtained with the longitudinal power (blue) and the cross power (green). The white dot marks the fiducial model in the parameter space. Comparing this plot with figure 3.7 makes clear why the cross power spectrum is not the optimal statistic for measuring λ_J since the phase information is diluted and the degeneracy is not efficiently broken.

The red shaded regions in Figure 3.7 show the constraints in thermal parameter space resulting from our MCMC exploration of the phase angle likelihood (eqn. 3.12). The results are shown projected onto the $T_0 - \lambda_J$ and $\gamma - \lambda_J$ planes, where the third parameter (γ and T_0, respectively) has been marginalized over. The dark and light shaded regions show 65% and 96% confidence levels, respectively. The phase difference technique (red) yields essentially horizontal contours, which pinpoint the value of the Jeans scale, with minimal degeneracy with other thermal parameters. This is a direct consequence of the near independence of the phase angle PDF of T_0 and γ shown in Figure 3.4, and discussed in § 3.1.4. The physical explanation for this independence is that 1) the non-linear FGPA transformation in only weakly dependent on temperature 2) phase angles are invariant to the thermal broadening convolution. This truly remarkable result is the key finding of this work: phase angles of the Lyα forest flux provide direct constraints on the 3D smoothing of the IGM density independent of the other thermal parameters governing the IGM.

The blue shaded regions in Figure 3.7 show the corresponding parameter constraints for our MCMC of the longitudinal power spectrum likelihood (eqn. 3.15). Considering the longitudinal power spectrum alone, we find that significant degeneracies exist between λ_J, T_0 and γ, which confirms our qualitative discussion of these degeneracies in § 2.3.1 and illustrated in Figure 2.2. These degeneracy directions are easy to understand. The longitudinal power is mostly sensitive to thermal parameters via the location of the sharp small-scale cutoff in the power spectrum. This thermal cutoff is set by a combination of both 3D Jeans pressure smoothing and 1D thermal broadening. The thermal broadening

component is set by the temperature of the IGM at the characteristic overdensity probed by the forest, which is $\delta \approx 2$ at $z = 3$ [Becker et al., 2011]. One naturally expects a degeneracy between T_0 and γ, because it is actually the temperature at $T(\delta \approx 2)$ that sets the thermal broadening. A degeneracy between λ_J and $T(\delta \approx 2)$ is also expected because both smoothings contribute to the small-scale cutoff. Thus, a lower Jeans scale can be compensated by more thermal broadening, which can result from either a steeper temperature density relation (larger γ) or a hotter temperature at mean density T_0, since both produce a hotter $T(\delta \approx 2)$.

Previous work that has aimed to measure thermal parameters such as T_0 and γ, from the longitudinal power spectrum [Viel et al., 2009, Zaldarriaga et al., 2001], the curvature statistic [Becker et al., 2011], wavelets [Garzilli et al., 2012, Lidz et al., 2009, Theuns et al., 2002b], and the b-parameter distribution [Bryan & Machacek, 2000, Haehnelt & Steinmetz, 1998, McDonald et al., 2001, Ricotti et al., 2000, Rudie et al., 2012, Schaye et al., 2000, Theuns et al., 2000, 2002a], have for the most part ignored the degeneracies between these thermal parameters and the Jeans scale (but see Zaldarriaga et al. 2001 who marginalized over the Jeans scale, and Becker et al. 2011 who also considered its impact). Neglecting the possible variation of the Jeans scale is equivalent to severely restricting the family of possible IGM thermal histories. Because the phase angle method accurately pinpoints the Jeans scale independent of the other parameters, it breaks the degeneracies inherent to the longitudinal power spectrum and will enable accurate and unbiased measurements of both T_0 and γ, as evidence by the intersection of the red and black contours in Figure 3.7. Similar degeneracies between the Jeans scale and (T_0,γ) exist when one considers other statistics such as the flux PDF [Bolton et al., 2008, Calura et al., 2012, Garzilli et al., 2012, Kim et al., 2007, McDonald et al., 2000], which we will explore in an upcoming study (Rorai et al., in prep). In light of these significant degeneracies with the Jeans scale, it may be necessary to reassess the reliability and statistical significance of previous measurements of T_0 and γ.

Figure 3.8 shows the resulting thermal parameter constraints for our MCMC analysis of the cross-power spectrum likelihood (eqn. 3.16) in green, determined from exactly the same mock quasar pair sample that we analyzed for the phase angles. The confidence regions for the longitudinal power are shown for comparison in blue. The cross-power spectrum is a straightforward statistic to measure and fit models to, and the green confidence regions clearly illustrate that it does exhibit some sensitivity to the Jeans scale, as discussed in § 2.3.2 and shown in the right panel of Figure 2.2. However, a comparison of the cross-power confidence regions in Figure 3.8 (green) with the phase angle PDF confidence regions in Figure 3.7 (red) reveals that there is far more information about the Jeans scale in quasar pair spectra than can be measured with the cross-power. The cross-power produces constraints which are effectively a hybrid between the horizontal

Jeans scale contours for the phase angle distribution and the diagonal banana shaped contours produced by the longitudinal power, which reflects the degeneracy between Jeans smoothing and thermal broadening. This quantitatively confirms our argument in § 3.1.1, that the cross-power is a product of moduli, containing information about the 1D power, and the cosine of the phase, which depends on the 3D power.

The results of this section indicate that among the statistics that we have considered, the phase angle PDF is the most powerful for constraining the IGM pressure smoothing, because it is more sensitive to the Jeans scale and results in constraints that are free of degeneracies with other thermal parameters. We demonstrate this explicitly in Figure 3.9, where we show the fully marginalized posterior distribution (see eqn. 3.13) of the Jeans scale for each the statistics we have considered. The probability distributions quantify the visual impression from the contours in Figures 3.7 and 3.8, and clearly indicate that the phase angle PDF is the most sensitive. The relative error on the Jeans scale $\sigma_\lambda/\lambda_J = 3.9\%$, which is a remarkable precision when compared to the typical precision $\sim 30\%$ of measurements of T_0 and γ in the published literature [see e.g. Figure 30 in Lidz et al., 2009, for a recent compilation], especially when one considers that only 20 quasar pair spectra are required to achieve this accuracy.

We close this section with a caveat to our statements that our Jeans scale constraints are free of degeneracies with other thermal parameters. The phase angle PDF is *explicitly* nearly independent of the temperature-density relation because 1) the non-linear FGPA transformation is only weakly dependent on temperature and 2) the phase angle PDF is invariant to the thermal broadening convolution (see § 3.1.4). However, in our idealized dark-matter only simulations, the Jeans scale is taken to be completely independent of T_0 and γ; whereas, in reality all three parameters are correlated by the underlying thermal history of the Universe. In this regard, the Jeans scale may *implicitly* depend on the T_0 and γ at the redshift of the sample, as well as with their values at earlier times. We argued that because the thermal history is not known, taking the Jeans scale to be free parameter is reasonable. However, the validity of this assumption and the implicit dependence of the Jeans scale on other thermal parameters is clearly something that should be explored in the future with hydrodynamical simulations.

3.2.4 The Impact of Noise and Finite Spectral Resolution

Up until this point we have assumed perfect data with infinite signal-to-noise ratio and resolution. This is unrealistic, especially considering, as discussed in § 3.2.2, that that close quasar pairs are faint, and typically cannot be observed at echelle resolution or very high signal-to-noise ratio $\gtrsim 20$, even with 8m class telescopes. In this section we

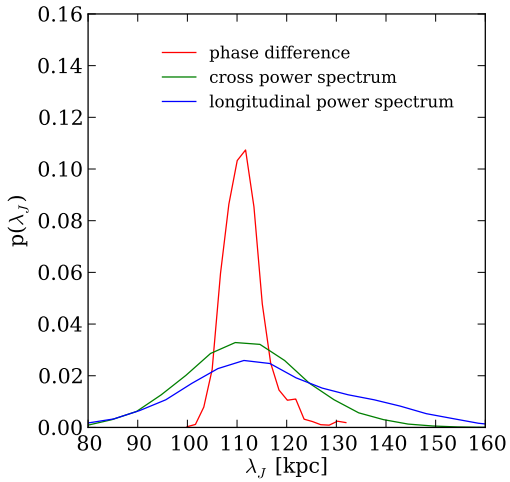

FIGURE 3.9: Estimated accuracy on the measurement of λ_J, obtained marginalizing over T_0 and γ the posterior distribution from the MCMC analysis. The phase difference statistic (red) sets tighter constraints than the cross power (blue) and the longitudinal power (black), which are affected by parameter degeneracies. In this case we do not account for the effect of noise and limited resolution, and we find a relative precision of 3.9% for λ_J.

explore the impact of noise and finite resolution on the precision with which we can measure the Jeans scale.

We consider the exact same sample of 20 mock quasar spectra, but now assume that they are observed with spectral resolution corresponding to FWHM $= 30\,\mathrm{km\ s^{-1}}$, and two different signal-to-noise ratios of S/N $\simeq 5$ and S/N $\simeq 10$ *per pixel*. These values are consistent with what could be achieved using an echellette spectrometer on an 8m class telescope. To create mock observed spectra with these properties, we first smooth our simulated spectra with a Gaussian kernel to model the limited spectral resolution, and interpolate these smoothed spectra onto a coarser spectral grid which has $10\,\mathrm{km\ s^{-1}}$ pixels, consistent with the spectral pixel scale of typical echellette spectrometers. We then add Gaussian white noise to each pixel with variance σ_{N}^2 determined by the relation $\mathrm{S/N} = \bar{F}/\sigma_{\mathrm{N}}$, where $\bar{F}$ is the mean transmitted flux. This then gives an average signal-to-noise ratio equal to the desired value.

As we already discussed in § 3.1.4 in the context of thermal broadening, phase angles are invariant under a convolution with a symmetric Gaussian kernel. Thus we do not expect spectral resolution to significantly influence our results, provided that we restrict attention to modes which are marginally resolved, such that we can measure their phases. Indeed, the cutoff in the flux power spectrum induced by spectral resolution is $k_{res} = 1/\sigma_{\mathrm{res}} \approx 2.358/\mathrm{FWHM} = 0.08\,\mathrm{s\ km^{-1}}$, is comparable to the maximum wavenumber we consider $k = 0.1\,\mathrm{s\ km^{-1}}$, and hence we satisfy this criteria. Note further that this invariance to a symmetric spectral convolution implies that we do not need to be able

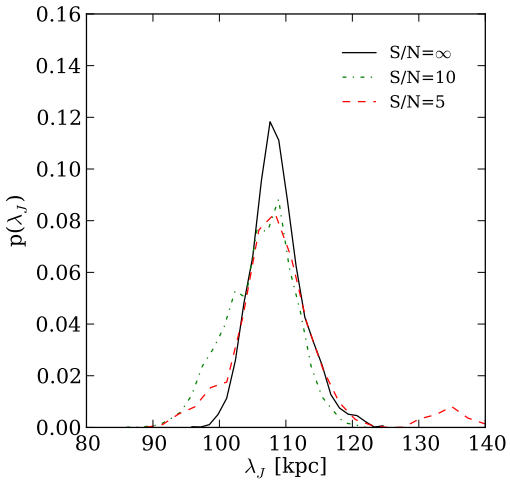

FIGURE 3.10: The effect of noise and resolution in the measurement of λ_J. The plots shows the posterior distribution of the Jeans scale, marginalized over T_0 and γ. Each line represent a different degree of noise, assuming a resolution of FWHM=30 km/s. We selected a different subsample of the simulation as our mock dataset which has a precision of 3.6% for S/N=∞ (black solid), 4.8% for S/N=10 (green dot-dashed) and 7.2% for S/N=5 (red dashed).

to precisely model the resolution, provided that it has a nearly symmetric shape and does not vary dramatically across the spectrum. This is another significant advantage of the phase angle approach, since the resolution of a spectrometer often depends on the variable seeing, and can be challenging to accurately calibrate.

Although our results are thus likely to be very independent of resolution, noise introduces fluctuations which are uncorrelated between the two sightlines, and this will tend to reduce the coherence of the flux that the phase angle PDF quantifies. Noise will thus modify the shape of the phase angle PDF away from the intrinsic shape shown in Figure 3.3. In order to deal with noise and its confluence with spectral resolution, we adopt a forward-modeling approach. Specifically, for each thermal model we smooth all 10,000 IGM skewers to finite resolution, interpolate onto coarser spectral grids, and add noise consistent with our desired signal-to-noise ratio. We then fit the resulting distribution of phase angles to the wrapped-Cauchy distribution, determining the value of the concentration parameter $\zeta(k, r_\perp)$, at each k and $r_\perp$ as we did before. We again emulate the function $\zeta(k, r_\perp|T_0, \gamma, \lambda_J)$ using the same thermal parameter grid, but now with noise and spectral resolution included, enabling fast evaluations of the likelihood in eqn. (3.12). Thermal parameter constraints then follow from MCMC exploration of this new likelihood, for which the impact of noise and resolution on the phase angle PDF have been fully taken into account.

In Figure 3.10 we show the impact of noise on the fully marginalized constraints on the Jeans scale from the phase angle PDF. The solid curve represents the posterior

distribution for a mock dataset with infinite resolution and signal-to-noise ratio, which is identical to the red curve in Figure 3.9. The dotted and dashed curves illustrate the impact of S/N = 10 and S/N = 5, respectively. Note that the slight shift in the modes of these distributions from the fiducial value are expected, and should not be interpreted as a bias. Different noise realizations generate scatter in the phase angles just like the intrinsic noise from large-scale structure. The inferred Jeans scale for any given mock dataset or noise realization will not be exactly equal to the true value, but should rather be distributed about it with a scatter given by the width of the resulting posterior distributions. The relative shifts in the mode of the posterior PDFs are well within 1σ of the fiducial value, and are thus consistent with our expectations.

The upshot of Figure 3.10 is that noise and limited spectral resolution do not have a significant impact on our ability to measure the Jeans scale. For a signal-to-noise ratio of S/N = 10 per pixel we find that the relative precision with which we can measure the Jeans scale is $\sigma_\lambda/\lambda_J = 4.8\%$, which is only a slight degradation from the precision achievable from the same dataset at infinite signal-to-noise ratio and resolution $\sigma_\lambda/\lambda_J = 3.9\%$. The small impact of noise on the Jeans scale precision is not surprising. For the 10 km s^{-1} spectral pixels that we simulate, the standard deviation of the normalized Lyα forest flux per pixel is $\sqrt{\langle \delta F^2 \rangle} \simeq 32\%$, whereas for S/N = 10 our Gaussian noise fluctuations are at a significantly smaller $\simeq 10\%$ level. Heuristically, these two 'noise' sources add in quadrature, and thus the primary source of 'noise' in measuring the phase angle PDF results from the Lyα forest itself, rather than from noise in the data. For a lower signal-to-noise ratio of S/N = 5 per pixel, the precision is further degraded to $\sigma_\lambda/\lambda_J = 7.2\%$, which reflects the fact that noise fluctuations are becoming more comparable to the intrinsic Lyα forest fluctuations.

These numbers on the scaling of our precision with signal-to-noise ratio S/N provide intuition about the optimal observing strategy. For a given sample of pairs, it will require four times more exposure time to increase the signal-to-noise ratio from S/N $\simeq 5$ to S/N $\simeq 10$, whereas the same telescope time allocation could be used to increase the sample size by a factor of four at the same signal-to-noise ratio (assuming sufficient close pair sightlines exist). For the latter case of an enlarged sample, the precision will scale roughly as $\propto \sqrt{N_{\rm pairs}}$, implying a $\sigma_\lambda/\lambda_J = 3.6\%$ for a sample of 80 pairs observed at S/N = 5. This can be compared to $\sigma_\lambda/\lambda_J = 4.8\%$ for 20 pairs observed at S/N $\simeq 10$. There is thus a marginal gain in working at lower S/N $\simeq 5$ and observing a larger pair sample, although we have not considered various systematic errors which could impact our measurement. However, higher signal-to-noise spectra are usually preferable for the purposes of mitigating systematics, and hence one would probably opt for higher signal-to-noise ratio, a smaller pair sample, and tolerate slightly higher statistical errors.

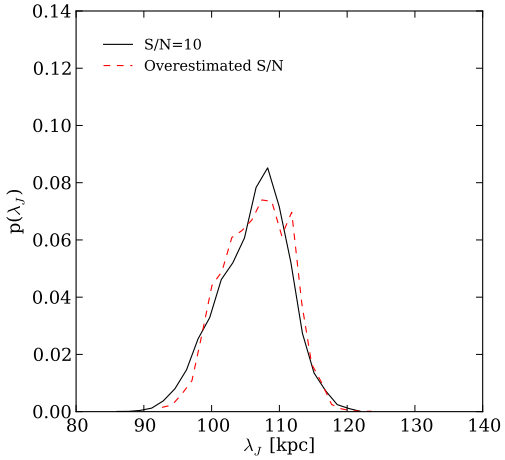

FIGURE 3.11: The effect of overestimating the signal-to-noise ratio by a 20% factor (red, dashed line) when the real value is S/N= 10: we do not find any significant bias on the measured value of the Jeans scale.

3.2.5 Systematic Errors

We now briefly discuss the systematic errors which could impact a measurement of the Jeans scale. First, consider the impact of errors in the continuum normalization. Because the phase angle is a ratio of Fourier modes of the normalized flux eqn. (3.3), it is completely insensitive to the continuum normalization of δF, provided that the continuum is not adding significant power on the scale of wavelength of the k-mode considered. In the previous section, we argued that finite spectral resolution does not have a significant impact the phase angle PDF, because phase angles are invariant under convolutions with symmetric kernels. We do take resolution into account in our forward-modeling of the phase angle PDF, but precise knowledge of the spectral resolution or the line spread function is not required, since the line spread function will surely be symmetric when averaged over several exposures, thus leaving the phase angles invariant. The only requirement is that we restrict attention to modes less than the resolution cutoff $k \lesssim k_{\rm res}$ whose amplitudes are not significantly attenuated, such that we can actually measure their phase angles.

Noise does modify the phase angle PDF, but our forward-modeling approach takes this fully into account provided the noise estimates are correct. One potential systematic is uncertainty in the noise model. The typical situation is that the standard-deviation of a spectrum reported by a data reduction pipeline is underestimated at the $\sim 10-20\%$ level (S/N overestimated), because of systematic errors related to the instrument and data reduction [see e.g. Lee et al., 2013, McDonald et al., 2006]. To address this issue we directly model the impact of underestimated noise for a case where we think the S/N $\simeq$ 10, but where in reality it is actually 20% lower S/N $\simeq$ 8. Specifically, using

our same mock dataset we generate 20 quasar pair spectra with S/N $\simeq$ 8. However, when forward-modeling the phase angle PDF with the IGM simulations, we take the signal-to-noise ratio to be the overestimated value of S/N $\simeq$ 10. Excess noise above our expectation would tend to reduce the coherence in the spectra (less peaked phase angle PDF) mimicking the effect of a smaller Jeans scale. We thus expect a bias in the Jeans scale to result from the underestimated noise. Figure 3.11 compares the posterior distributions of the Jeans scale for the two cases S/N $\simeq$ 10 (black curve) and signal-to-noise ratio overestimated to be S/N $\simeq$ 10 but actually equal to S/N $\simeq$ 8 (red curve). We see that $\simeq$ 20% level uncertainties in the noise lead to a negligible bias in the Jeans scale.

The only remaining systematic that could impact the Jeans scale measurement is metal-line absorption within the forest. Metal absorbers cluster differently from the IGM, and it is well known that metals add high-k power to the Lyα forest power spectrum because the gas traced by metal lines tends to be colder than H I in the IGM [Croft et al., 2002, Kim et al., 2004, Lidz et al., 2009, McDonald et al., 2000]. As this metal absorption is not present in our IGM simulations, it can lead to discrepancies between model phase angle PDFs and the actual data, resulting in a biased measurement. This is very unlikely to be a significant effect. We restrict attention to large scale modes with $k < 0.1\,\mathrm{s\ km^{-1}}$, both because this is comparable to our expected spectral resolution cutoff, and because below these wavenumbers metal line absorption results in negligible contamination of the longitudinal power [Croft et al., 2002, Kim et al., 2004, Lidz et al., 2009, McDonald et al., 2000]. Since the metal absorbers have a negligible effect on the *moduli* of these large scale modes, we also expect them to negligibly change their phase angles.

We thus conclude that the phase angle PDF is highly insensitive to the systematics that typically plague Lyα forest measurements, such as continuum fitting errors, lack of knowledge of spectral resolution, poorly calibrated noise, and metal line absorption.

3.2.6 Is Our Likelihood Estimator Unbiased?

Finally, we determine whether our procedure for measuring the Jeans scale via the phase angle likelihood (eqn. 3.12) outlined at the end of § 2.2, produces unbiased estimates. To quantify any bias in our Jeans scale estimator we follow a Monte Carlo approach, and generate 400 distinct quasar pair samples by randomly drawing 20 quasar pair spectra (allowing for repetition) from our ensemble of 10,000 skewers. Note that the distribution of transverse separations is approximately the same for all of these realizations, since we only simulate 30 discrete separations, and the full sample of 20 overlapping pair spectra requires 200 pairs of skewers, which are randomly selected from among the 30 available

pair separations. We MCMC sample the likelihood in eqn. (3.12) for each realization, and thus generate the full marginalized posterior distribution (eqn. 3.13; red curve in Figure 3.9). The 'measured' value of the Jeans scale for each realization is taken to the be the mean of the posterior distribution. We conducted this procedure for the case of finite spectral resolution (FWHM $= 30\,\mathrm{km\ s^{-1}}$) and signal-to-noise ratio S/N $\simeq 5$, where our forward-modeling procedure described in § 3.2.4 is used to model the impact of resolution and noise on the phase angle PDF.

The distribution of Jeans scale measurements resulting from this Monte Carlo simulation is shown in Figure 3.12. We find that the distribution of 'measurements' is well centered on the true value of $\lambda_J = 110$ kpc, and the mean value of this distribution is $\lambda_J = 111.1$ kpc, which differs from the true value by only 1%, confirming that our procedure is unbiased to a very high level of precision. The relative error of our measurements from this Monte Carlo simulation is $\sigma_{\lambda_J}/\lambda_J = 6.3\%$, which is consistent with the value of $\sigma_{\lambda_J}/\lambda_J = 7.2\%$, which we deduced in § 3.2.3 from an MCMC sampling of the likelihood for a single mock dataset. This confirms that the posterior distributions derived from our MCMC do indeed provide an accurate representation of the errors on the Jeans scale and other thermal parameters. However, we note that there is some small variation in the value of $\sigma_{\lambda_J}/\lambda_J$ inferred from the posterior distributions for different mock data realizations, as expected. Given that we only generated 400 samples, the error on our determination of the mean of the distribution in Figure 3.12 is $\simeq \sigma_{\lambda_J}/\lambda_J/\sqrt{400} = 0.3\%$, and thus our slight bias of 1% constitutes a $\sim 3\sigma$ fluctuation. We suspect that this is too large to be a statistical fluke, and speculate that a tiny amount of bias could be resulting from interpolation errors in our emulation of the IGM. It is also possible that choosing an alternative statistic of the posterior distribution as our 'measurement' instead of the mean, for example the mode or median, could also further reduce the bias. But we do not consider this issue further, since the bias is so small compared to our expected precision.

We conclude that our phase angle PDF likelihood procedure for estimating the Jeans scale has a negligible $\simeq 1\%$ bias. We would need to analyze a sample of $\simeq 500 - 1000$ quasar pair spectra for this bias to be comparable to the error on the Jeans scale. Furthermore, it is likely that we could, if necessary, reduce this bias even further by either reducing the interpolation error in our emulator or by applying a different statistic to our posterior distribution to determine the measured value.

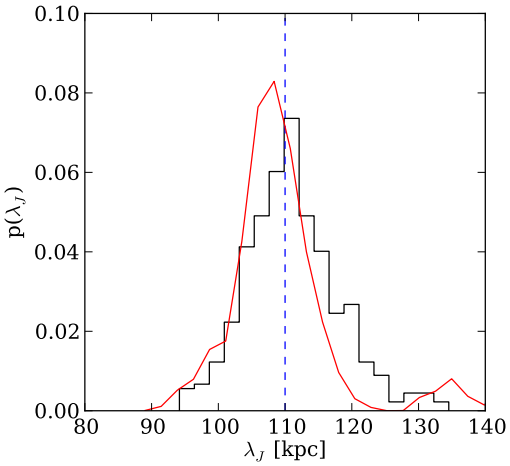

FIGURE 3.12: Probability distribution of the measured value of λ_J for 400 different mock datasets drawn from the fiducial simulation. This plot confirms that our method is not biased, since the distributions is be centered at the true value, marked with a vertical dashed line. This test is performed assuming S/N= 5. The red line is the posterior distribution deduced from our MCMC sampling of the phase angle PDF likelihood for one of these 400 mock dataset realizations. Its similarity in shape to the distribution of mock measurements illustrates that our MCMC simulations provide reliable error estimates.

Chapter 4

Data Analysis

In the previous chapter, I described how a sample of quasar pairs at separations in the range $\approx 50 - 500$ ckpc can be used to constrain the filtering scale down to precision of few percents. This prediction is based on a set of semi-analytical models based on a N-body dark matter simulation. The next goal is to do the same analysis on a sample of observed quasar pairs in order to measure λ_J at different redshifts and provide a rigorous estimation of its uncertainty.

To accomplish this task we have to face two main challenges:

- the formalism described in the previous chapters must be generalized to include a consistent treatment of noise, resolution and other possible systematics present in the data;
- we need to proof that our set of simplified IGM models, that do not include full hydrodynamic, are accurate enough to yields meaningful results.

The present chapter focuses on how we address the first problem, while a discussion on the second one is deferred to 6.

Although in § 3.2.4 we have discussed the effect of noise on phase distributions, this have been done in a very simplified manner, by assuming a constant noise and resolution across the whole box. Such analysis is not suitable for the sample of observed pairs that we want to analyze. Their spectra have been collected using different instruments and under different conditions, and therefore they have a wide range of resolutions and signal-to-noise ratios (S/N), which are also wavelength-dependent. This diversity motivates a specific calibration of phase differences for each single pair.

This chapter is structured as follows: in section 4.1 we briefly illustrate the pair sample that we use and we specify the requirement we set on data. We describe in section

4.2 the method we adopt to calculate phases from the spectra and in § 4.3 how the simulations are calibrated to the data sample in order to produce predictions for the PDFs of observed phases.

4.1 Data sample

This project is based on a large sample of quasar pairs drawn mostly from the SDSS [Abazajian et al., 2009] and BOSS[Ahn et al., 2012] and (in few cases) from the 2QZ[Croom et al., 2004] surveys. Beside studying the transverse coherence of the Lyα forest, such objects have been used for a number of scientific goals. Examples are the measurement of the small-scale clustering of quasars [Hennawi et al., 2006b, Myers et al., 2008, Shen et al., 2010] and the characterization of the circumgalactic medium of quasar hosts [Hennawi & Prochaska, 2007, 2008, Hennawi et al., 2006a, Prochaska & Hennawi, 2009, Prochaska et al., 2013, 2012].

Surveys such as SDSS and BOSS select *against* small separations of quasar pairs due to fiber collision, setting a lower limit to the angular separation of $55'', 62''$ and $30''$ for the SDSS, BOSS and the 2QZ survey, respectively. These angles are unfortunately too wide to probe the Jeans scale, thus follow-up spectroscopy is necessary in order both to discover close companions around quasars and to obtain science-quality spectroscopic data. An alternative possibility to find close pairs is to use the SDSS five-band photometry to select candidate companions around known quasars. A number of such candidates have been spectroscopically confirmed using the Apache Point Observatory (APO)[Hennawi et al., 2006b].

4.1.1 Spectroscopic Observations

A significant fraction of our dataset rely on spectra from SDSS and BOSS which have a resolution of $R \approx 2000$ and wavelength coverage of $\lambda \approx 3800 - 9000$ Åand $\lambda \approx 3600 - 10000$ Å, respectively. The rest of the spectra have been collected with follow-up spectroscopy on large-aperture telescopes, using instruments with a wide range of capabilities which I list below.

Part of objects were observed at the Keck 10m telescope, including data from the Echellette Spectrometer and Imager (ESI, Sheinis+2002), the Low Resolution Imaging Spectrograph (LRIS; oke et al 1995), and the High Resolution Echelle Spectrometer (HIRES, Vogt+94). Other spectra were collected using the Gemini MultiObject Spectrograph (GMOS, hook+2004) on the 8m Gemini North and South telescopes, the

Magellan Echellette Spectrograph (MagE, Marshal+08) and the Magellan Inamori Kyocera Echelle (MIKE, Bernstein+2003) on the 6m Magellan telescopes. Few pairs were observed through the Multi-Object Double Spectrograph (MODS, Pogge+2012) on the Large Binocular Telescope (LBT). We recently obtained new data from the X-Shooter spectrometer at the Very Large Telescope (VLT), specifically selected for this project. The GMOS spectra have been observed in the context of two different programs, in which gratings of 600 lines/mm (GMOS600) and 1200 lines/mm (GMOS1200)were used.

An exhaustive description of the properties of the data sample can be found in Hennawi & Prochaska [2008], Hennawi et al. [2006a], Prochaska et al. [2013].

4.1.2 Selection Criteria

We apply a first broad cut to select pairs suitable for the science goal of this study. An obvious requisite is the existence of a segment of coeval Lyα forest, which can be expressed as $(1+z_{\rm fg})\lambda_{{\rm Ly}\alpha} > (1+z_{\rm bg})\lambda_{{\rm Ly}\beta}$. However, we want to avoid cases where this segment is too small to calculate meaningful statistics, especially considering that we cannot use the wavelengths too close to the Lyα and Lyβ emission lines. We define the "overlapping fraction" of the Lyα forest as

$$f_{\rm ov} = \frac{(1+z_{\rm fg})\lambda_{{\rm Ly}\alpha} - (1+z_{\rm bg})\lambda_{{\rm Ly}\alpha}}{(1+z_{\rm fg})(\lambda_{{\rm Ly}\alpha} - \lambda_{{\rm Ly}\beta})} \tag{4.1}$$

and we set a lower threshold at $f_{\rm ov} = 0.3$, removing in this way part of the projected pairs with the highest redshift separation.

A second criterion is established according to the transverse separation. Our study of the sensitivity of phases with simulations (see chapter 3) indicated that the most informative pairs are those with impact parameter comparable to the Jeans scale. Considered that the line-of-sight power of the forest excludes Jeans scales larger than $\approx 300-400kpc$, we focus our analysis on pairs closer than 500 kpc (comoving) at the f/g redshift. Only at redshift $z > 3$ we loose this restriction up to 700 kpc, since the sample at this redshift is considerably smaller and even the weak constraints coming from wide pairs are valuable.

We than exclude from the sample all pairs for which no science-quality spectra is available because they have not been observed with one of the instrument listed in the previous paragraph. Future programs of follow-up spectroscopy will allow this objects to be used in the measurement.

The set of pairs selected at the end of this process is then visually inspected in order to find contaminants. Some of the QSOs exhibit strong associated absorption lines known as Broad Absorption Lines (BAL), which are thought to be produced in the vicinity of

the black hole and may reach velocities up to $v \gtrsim 10000$ km/s. For this reason they could be blueshifted into the Lyα forest, causing blending with IGM absorption. Since we are not able to model this blending, we remove from the sample all the pairs in which one of the two spectrum is contaminated by BAL. DLAs and LLSs may also pose problems, since they fall out of the optically thin approximation where our model is valid. Therefore we isolate and mask those regions of the spectra where we can identify such absorbers. This is unfortunately not easy to do with LLSs, however we reckon that their impact on phase difference should be small, given that their contribute to the Lyα-forest absorprtion is very small [McDonald et al., 2005].

We decide also to exclude all pairs which are known to be lenses, i.e. they are just a double image of the same source. In principle they can be used if the lens redshift is precisely known, in which case the dependence of the impact parameter with redshift could be easily modeled. However, this generally leads to very small separations at Lyα forest redshift ($\lesssim 10$ kpc), which might be too tiny to probe the Jeans scale and sensitive to the physics of very small scales which we are not capturing in our models. For this reasons, we leave the transverse analysis of lenses to future projects.

We do not set tight limits on the signal-to-noise ratio, since we try to model noise and we expect to detect signal from large-scale modes also for noisy data. We preliminary demand S/N $\gtrsim 5$ for 1-Åpixels or equivalent, but we apply a further cut based on a test we perform *a posteriori* which will be described in § 5.1.5.

The final sample obtained through this procedure is illustrated in figure 4.1, where each pair is depicted as a black line. The lines trace the coeval forest in pairs, following the evolution of the impact parameter as a function of redshift. The extension of the overlapping segments depends on redshift, on $f_{\rm ov}$ and on the removals of contaminants, which appear as 'holes' in the lines. The vertical red dashed lines delimit the three redshift bins on which we perform our analysis $[1.8, 2.2], [2.2, 2.7]$ and $[2.7, 3.3]$. The lower limit $z = 1.8$ is set to avoid the forest close to the atmospheric cutoff, and the bins are wider at higher z to enclose a sufficiently large sample of pairs. New observations are required to extend the measurement to $z > 3.3$. A complete list of the coeval Lyα-forest chunks is provided in table 4.1, together with all the relevant parameters. Note that the forest of a quasar pair may be split in more than a chunk if we need to remove a segment due to contaminants.

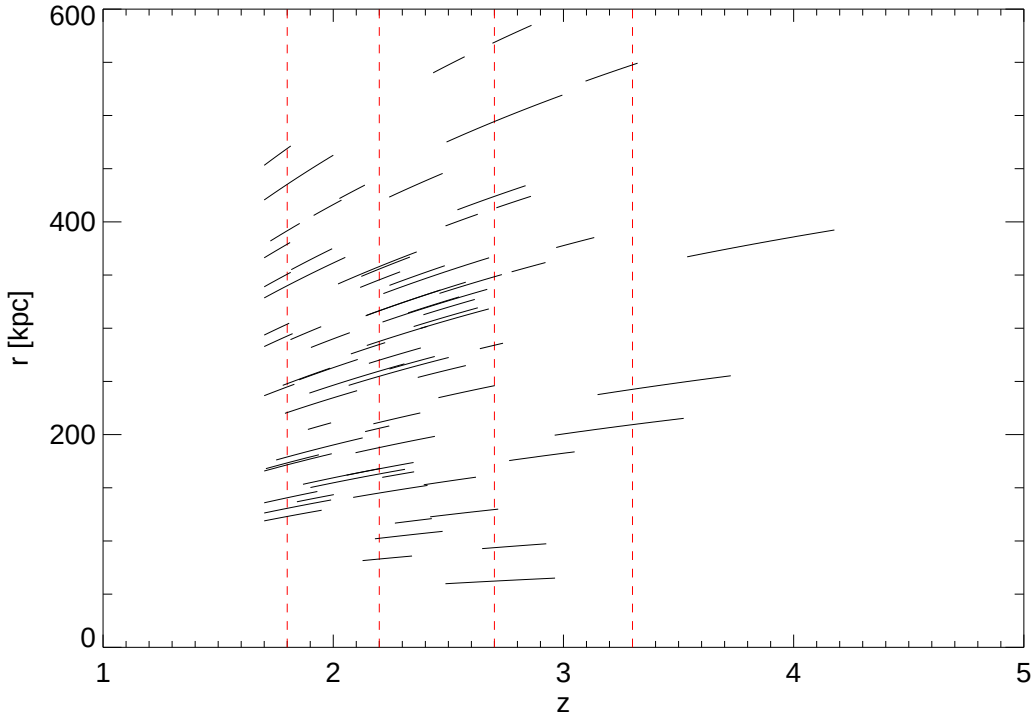

FIGURE 4.1: The distribution of our sample in redshift z and in transverse separation $r_\perp$. Each line represents a segment of overlapping Lyα forest in a pair. The length of a segment depends on the redshifts of the two quasars and on the presence of DLAs or other contaminants that require to exclude part of the forest. The atmospheric cutoff sets a lower limit for all pairs at $z \approx 1.7$. The lines are curved because the impact parameter evolves with redshift converging toward us. The four red dotted lines delimit the three redshift intervals in which we split the sample.

TABLE 4.1: Complete list of the chunks of overlapping Lyα forest in the pair sample we analyze

Name	$z_{\rm bg}$[a]	$z_{\rm fg}$[b]	$z_{\rm min}$[c]	$z_{\rm max}$[d]	θ[e]	$r_\perp$[f]	Instrument	$R_{\rm bg/fg}$[g]	$S/N_{\rm bg/fg}$[h]
SDSSJ0034-1049	1.95	1.83	1.61	1.77	7.6	177	LRIS/LRIS	180/180	6.9/33.7
SDSSJ0054-0946	2.12	2.12	1.67	2.05	14.1	347	LRIS/LRIS	174/174	133.5/28.0
SDSSJ0117+3153	2.64	2.62	2.33	2.55	11.3	322	ESI/ESI	64/64	30.9/20.0
SDSSJ0214+0105	2.29	2.21	2.03	2.14	16.4	428	MODS/MODS	205/205	9.6/6.0
SDSSJ0332-0722	2.11	2.10	1.66	2.00	18.1	440	LRIS/LRIS	175/175	16.5/17.4
SDSSJ0735+2957	2.08	2.06	1.63	1.99	5.4	131	LRIS/LRIS	176/176	53.7/34.6
SDSSJ0750+2724	1.80	1.77	1.60	1.71	13.1	299	LRIS/LRIS	182/182	6.6/14.3
SDSSJ0752+4011	2.12	1.87	1.67	1.81	12.6	298	LRIS/LRIS	178/178	15.3/6.3
SDSSJ0813+1014	2.08	2.06	1.65	1.99	7.1	173	LRIS/LRIS	176/176	18.3/15.0
SDSSJ0814+3250	2.21	2.17	1.85	2.11	10.3	261	GMOS1200/GMOS1200	191/191	6.5/12.7
SDSSJ0837+3837	2.25	2.05	1.78	1.99	10.3	255	LRIS/LRIS	174/174	8.0/37.5
SDSSJ0853-0011	2.58	2.41	2.12	2.33	13.2	358	MAGE/MAGE	62/62	38.3/8.6
SDSSJ0913-0107	2.92	2.75	2.35	2.63	10.8	311	GMOS600/GMOS600	192/192	9.3/11.7
SDSSJ0920+1310	2.43	2.42	2.06	2.35	6.2	168	MAGE/MAGE	62/62	25.8/37.7
SDSSJ0924+3929	2.08	1.88	1.64	1.82	12.2	286	LRIS/LRIS	180/180	13.1/16.1
SDSSJ0937+1509	2.55	2.54	2.14	2.47	11.7	324	GMOS600/GMOS600	201/201	6.4/11.1
SDSSJ0938+5317	2.32	2.07	1.84	2.00	5.6	140	LRIS/LRIS	171/171	23.0/6.2
SDSSJ0956+2643	3.08	3.08	2.49	3.00	16.5	498	ESI/ESI	64/64	16.4/25.1
SDSSJ1006+4804	2.60	2.30	2.08	2.23	10.6	281	LRIS/LRIS	161/161	22.7/12.2
SDSSJ1009+2500	1.99	1.88	1.64	1.82	14.6	342	LRIS/LRIS	180/180	16.7/24.2
SDSSJ1021+1112	3.85	3.83	3.15	3.73	7.4	247	ESI/ESI	64/64	56.5/26.2
SDSSJ1026+0629	3.12	2.89	2.64	2.74	9.5	283	MAGE/MAGE	62/62	6.1/6.0
SDSSJ1053+5001	3.08	3.05	2.49	2.96	2.1	63	ESI/ESI	64/64	8.4/9.6
2QZJ1056-0059	2.13	2.12	1.71	1.94	7.2	175	LRIS/LRIS	177/177	13.2/10.5
SDSSJ1116+4118	3.00	2.94	2.49	2.63	13.8	402	LRIS/LRIS	130/130	22.4/45.9
SDSSJ1116+4118	3.00	2.94	2.71	2.86	13.8	419	LRIS/LRIS	123/123	26.0/52.9

Name	$z_{\rm bg}^{\rm a}$	$z_{\rm fg}^{\rm b}$	$z_{\rm min}^{\rm c}$	$z_{\rm max}^{\rm d}$	$\theta^{\rm e}$	$r_{\perp}^{\rm f}$	Instrument	$R_{\rm bg/fg}^{\rm h}$	$S/N_{\rm bg/fg}^{\rm i}$
SDSSJ1135-0221	3.02	3.01	2.77	2.92	11.6	357	GMOS600/GMOS600	175/175	6.0/6.0
SDSSJ1141+0724	3.79	3.55	3.10	3.32	16.7	541	GMOS600/GMOS600	160/160	9.6/7.8
SDSSJ1150+0453	2.52	2.52	2.10	2.44	7.0	191	GMOS600/GMOS600	204/204	7.2/9.6
SDSSJ1204+0221	2.53	2.44	2.02	2.36	13.3	357	MAGE/MAGE	62/62	19.2/28.5
SDSSJ1225+5644	2.39	2.38	1.90	2.31	6.1	159	LRIS/LRIS	163/163	13.2/34.0
SDSSJ1240+4329	3.26	3.25	2.65	2.93	3.1	95	GMOS600/GMOS600	177/177	6.6/11.6
SDSSJ1306+6158	2.17	2.10	1.73	1.85	16.3	390	LRIS/LRIS	177/177	6.1/5.8
SDSSJ1306+6158	2.17	2.10	1.91	2.04	16.3	413	LRIS/LRIS	169/169	9.6/11.1
SDSSJ1307+0422	3.04	3.01	2.46	2.70	8.2	241	MAGE/MIKE	62/8	33.0/29.7
SDSSJ1358+2737	2.11	1.89	1.66	1.83	10.2	241	LRIS/LRIS	179/174	10.5/23.7
SDSSJ1405+4447	2.22	2.20	1.75	2.13	7.4	187	LRIS/LRIS	172/172	13.8/52.6
SDSSJ1409+5225	2.11	1.88	1.69	1.82	19.5	462	LRIS/LRIS	177/177	21.9/6.0
SDSSJ1420+2831	4.31	4.29	3.54	4.18	10.9	380	ESI/ESI	64/64	17.8/14.6
SDSSJ1420+1603	2.06	2.01	1.81	1.95	12.0	295	MAGE/LRIS	62/169	9.0/14.6
SDSSJ1427-0121	2.35	2.27	1.87	2.20	6.2	161	MAGE/MAGE	62/62	25.1/20.0
SDSSJ1428+0232	3.02	3.01	2.43	2.57	19.0	548	GMOS600/GMOS600	191/191	10.0/9.6
SDSSJ1428+0232	3.02	3.01	2.69	2.86	19.0	576	GMOS600/GMOS600	178/178	11.0/11.5
SDSSJ1443+2008	2.67	2.65	2.14	2.58	11.7	328	SDSS/SDSS	150/150	6.9/10.1
SDSSJ1508+3635	2.10	1.84	1.65	1.78	15.2	356	LRIS/LRIS	179/179	8.1/29.9
SDSSJ1514+2101	2.24	2.19	1.79	2.10	9.2	231	MODS/MODS	215/215	6.1/5.2
SDSSJ1541+2702	3.63	3.62	2.96	3.52	6.4	208	ESI/ESI	64/64	9.5/13.9
SDSSJ1613+0808	2.39	2.38	1.90	2.31	9.6	253	MAGE/MAGE	62/62	31.9/18.3
SDSSJ1613+1616	2.76	2.76	2.22	2.68	12.3	350	GMOS600/GMOS600	194/194	12.7/11.6
SDSSJ1622+0702	3.26	3.23	2.76	3.05	5.8	180	ESI/ESI	64/64	115.9/18.0
SDSSJ1657+3105	2.39	2.14	1.90	2.07	11.3	289	MODS/MODS	212/211	10.1/18.2
SDSSJ1719+2549	2.17	2.17	1.82	2.00	14.7	365	GMOS1200/GMOS1200	196/196	9.4/9.5
SDSSJ2103+0646	2.57	2.55	2.18	2.48	3.8	106	GMOS600/GMOS600	200/200	7.6/8.8
SDSSJ2128-0617	2.07	2.06	1.89	1.99	8.3	208	LRIS/LRIS	170/170	35.0/9.7

Name	$z_{\rm bg}^{\rm a}$	$z_{\rm fg}^{\rm b}$	$z_{\rm min}^{\rm c}$	$z_{\rm max}^{\rm d}$	$\theta^{\rm e}$	$r_{\perp}^{\rm f}$	Instrument	$R_{\rm bg/fg}^{\rm h}$	$S/N_{\rm bg/fg}^{\rm i}$
SDSSJ2214+1326	2.01	2.00	1.57	1.93	5.8	138	LRIS/LRIS	179/179	29.1/31.8
SDSSJ2243-0613	2.59	2.58	2.07	2.50	9.5	260	GMOS600/GMOS600	203/203	7.1/12.6
SDSSJ2300+0155	2.95	2.91	2.38	2.68	10.7	309	MAGE/MAGE	62/62	11.9/21.6

[a]Redshifts of b/g quasar.
[b]Redshifts of f/g quasar.
[c]Minimum redshift of the chunk.
[d]Maximum redshift of the chunk.
[e]Angular separation between f/g and b/g quasar (arcsec).
[f]Impact parameter at f/g quasar redshift (comoving kpc).
[g]Mean resolution in the chunk (b/g-f/g).
[h]Mean signal-to-noise ration in the chunk (b/g-f/g).

4.1.3 Continuum Fitting and Data Preparation

We fitted the continuum manually for those pairs which were not already fitted for other projects. We used a fitting algorithm that perform a cubic spline interpolation between manually-inserted guiding points. We stress the fact that the statistic we use is not particularly sensitive to continuum-placement, as we will explicitly show in the next chapter. In particular, it is completely insensitive to its renormalization, while it could be affected by fluctuation on scales $\lesssim 2000$ km/s which we do not expect to find in quasar spectra. The noise has been estimated following the standard pipeline of the instruments.

We exclude the parts of spectrum close to the Ly$\alpha$and Lyman$-\beta$ emission lines, restricting the analysis to the rest-frame wavelength interval $[1040, 1190]$ nm. The overlapping forest in a pairs on which we calculate phases is thus defined by $\lambda \in [1040(1 + z_{\rm bg}), 1190(1 + z_{\rm fg})]$, which is slightly narrower than the one implied by the $f_{\rm ov}$ defined above.

Since phase differences are calculated in velocity space, we transform from wavelengths to velocities according to the formula

$$\Delta v = c \log(\lambda_1/\lambda_2), \tag{4.2}$$

where Δv is the relative velocity between two points responsible of resonant Lyα absorption at observed wavelengths λ_1 and λ_2. In the limit where peculiar velocities are negligible, this corresponds to a comoving distance of

$$\Delta x = \frac{(1 + z)\Delta v}{H(z)}. \tag{4.3}$$

where the $H(z)$ is the Hubble parameter at the observed redshift.

4.2 Calculation of Phases from Real Spectra

Applying Fourier-space statistics, such as phase difference, to the observed Lyα forest is not a straightforward operation. While in simulations we generate mock spectra on perfectly regular grids in velocity space, the pixels of observed spectra are in most of the cases unevenly distributed. Since the discrete Fourier transformation is defined for evenly-sampled functions, we have either to interpolate or rebin the data onto a regular grid, or to use approximate methods without modifying the sampling. The two methodologies have opposite advantages and disadvantages, so we decided to implement both and check that they lead to consistent results.

4.2.1 Method 1: Least-Square Spectral Analysis

A widely-used approach to generalize Fourier transformation to irregularly-sampled series is the so called least-square spectral analysis (LSSA). Practically speaking, it consists in fitting a function $f(x_i)$ with a linear combinations of trigonometric functions $\cos(k_j x_i)$ and $\sin(k_j x_i)$, where $\{x_i\}$ is the set of points where f is sampled and $\{k_j\}$ are the wavenumbers of the modes that we want to fit. This leads for example to the Lomb-Scargle periodogram [Lomb, 1976], a method often employed to calculate the power spectrum of a signal. It is also possible to follow this strategy to recover the phase information, which is what we want to calculate in quasar spectra. We follow for this purpose the method described in Mathias et al. [2004] which I briefly report here.

As mentioned above, the decomposition can be view as the minimization, *for each different* k_j of

$$||f(\mathbf{x}) - \mathbf{c_j}\cos(\mathbf{k_j x}) + \mathbf{s_j}\sin(\mathbf{k_j x})|| \tag{4.4}$$

where in our case $\mathbf{x}$ is the array of the velocity-space pixels in a spectrum, f is the transmitted flux of the Lyα forest, $||g(\mathbf{x})|| = \sum_\mathbf{i} \mathbf{g^2}(\mathbf{x_i})$ denotes the squared norm and $(c_j s_j)$ are the coefficients that we need to determine. Otherwise stated, we want to find the projection of f on the functional subspace defined by the linear combinations of $\cos(k_j\mathbf{x})$ and $\sin(k_j\mathbf{x})$. In the case where x_i are evenly spaced and $k_j = 2\pi j/L$, with L being the total length of the spectrum, this is equivalent Fourier decomposition. For generic $\{x_i\}$ and $\{k_j\}$ the linear subspaces relative to different k may not be orthogonal and may not form a complete functional base, so this fitting procedure cannot be properly regarded as a decomposition.

The minimization of expression 4.4 is obtained via the Moore-Penrose pseudo-inverse matrix [Penrose, 1955] applied to the linear system

$$f(\mathbf{x}) = (\mathbf{c_j}\ \mathbf{s_j})\mathbf{\Omega_j} \tag{4.5}$$

where Ω_j is defined as

$$\Omega_j = \begin{pmatrix} \cos(k_j x_1) & \dots & \sin(k_j x_n) \\ \sin(k_j x_1) & \dots & \sin(k_j x_n) \end{pmatrix}. \tag{4.6}$$

The pseudo inverse is then $\Omega_j^+ = \Omega_j^T(\Omega_j\Omega_j^T)^{-1}$ and the coefficients are estimated by

$$(c_j\ s_j) = f(\mathbf{x})\mathbf{\Omega_j^+}. \tag{4.7}$$

According to the pseudo-inverse properties, this coefficients are exactly the ones that minimizes $||f(\mathbf{x}) - (\mathbf{c_j}\ \mathbf{s_j})\mathbf{\Omega_j}||$, i.e. expression 4.4. When the system has a solution this

norm is zero, but for our problem this is never the case. Note also that this definition of the pseudo-inverse requires that $\Omega_j \Omega_j^T$ is invertible, which is however always satisfied for reasonable pixel distributions.

By writing explicitly eq. 4.7 we obtain

$$(c_j \; s_j) = \begin{pmatrix} \sum_i f(x_i)\cos(k_j x_i) \\ \sum_i f(x_i)\sin(k_j x_i) \end{pmatrix}^T \begin{pmatrix} \sum_i \cos^2(k_j x_i) & \sum_i \cos(k_j x_i)\sin(k_j x_i) \\ \sum_i \cos(k_j x_i)\sin(k_j x_i) & \sum_i \sin^2(k_j x_i) \end{pmatrix}^{-1} \tag{4.8}$$

where the diagonal terms are nonzero because $\sin(k_j \mathbf{x})$ and $\cos(k_j \mathbf{x})$ are not orthogonal in general. Nevertheless it is possible to apply a phase shift to the coordinates such that, for a given k_j, the non diagonal terms vanish [Lomb, 1976]. It can be shown that the shift is equal to

$$T_j = \frac{1}{2k} \arctan \frac{\sum_i \sin(k_j x_i)}{\sum_i \cos(k_j x_i)}. \tag{4.9}$$

After diagonalization, the equation above simplifies in

$$(c_j \; s_j) = \begin{pmatrix} \frac{\sum_i f(x_i)\cos(k_j(x_i - T_j))}{\sum_i \cos^2(k_j(x_i - T_j))} & \frac{\sum_i f(x_i)\sin(k_j(x_i - T_j))}{\sum_i \sin^2(k_j(x_i - T_j))} \end{pmatrix} \tag{4.10}$$

Which is the expression are looking for. The power spectrum immediately follow from this result as $P(k_j) = c_j^2 + s_j^2$.

If we need to recover phase information we must consider that phases are changed by the Lomb shift, therefore we have to apply at each k the inverse translation. This is easily done by defining the Fourier coefficients in the complex representation as

$$F(k_j) = (c_j + i s_j) e^{i k_j T_j}. \tag{4.11}$$

We are now ready to calculate phase differences in the usual way

$$\theta_{12}(k) = \arccos \left(\frac{\Re[\tilde{F}_1^*(k)\tilde{F}_2(k)]}{|\tilde{F}_1(k)||\tilde{F}_2(k)|} \right) \tag{4.12}$$

where F_1 and F_2 are the transmitted fluxes of the Lyα forest in the two spectra of the pair.

A final caveats concerns non-orthogonality: if the Fourier components have non zero cross products $C_{l,m} = \sum_i exp(-i(k_m - k_l)x_i)$, then the estimated Fourier coefficients are correlated. This would be an undesirable complication when calculating the likelihood function of phases, which we consider to be independent on the wake of or test in § 3.1.5. We solve this problem recursively: after calculating $\tilde{F}(k_0)$ we subtract this component

from the original function

$$F'(x) = F(x) - \tilde{F}(k_0)e^{-ik_0x} \tag{4.13}$$

and then we calculate the next coefficient $\tilde{F}(k_1)$ on the residual function $F'(x)$. We iterate this steps until all the coefficients are calculated. This algorithm is equivalent to a Grahm-Schmidt process, and requires to specify the order on which the components are subtracted. The most natural choice for us is starting with the large scale modes, i.e. with the lowest wavenumber, which are the least affected by noise and other systematics.

4.2.2 method 2: Rebinning on a Regular Grid

A second possibility is to rebin the observed flux pixels into a regular grid, to allow the standard calculation of the Fourier coefficients. The advantage of this method is that we avoid approximations deriving from the least-square evaluation of the phases, but on the other hand, we do not have a clear control on how the rebinning modifies the Fourier phases. The pros and cons of this approach are complementary to the LSSA procedure described in the previous section, therefore we decide to adopt both of them and check that the results are consistent, assuring in this way that rebinning or LSSA are not a source of bias (§ 5.3.1.1).

In order to consistently calculate phase differences, one need not only to bin the pixels of each spectrum in a regular grid in velocity space, but also to use the same regular grid for the two spectra of a pair. The common regular grid is defined from the original arrays via a simple procedure. for a single spectrum with N irregular pixels located at $\{v_i^0\}$, the step of the regularized array would be defined as $\Delta v = (v_N^0 - v_1^0)/N$ and the full vector would be $\{v_i = v_1^0 + i\Delta v\}$. When considering two spectra with different pixels arrays $\{u_i^0\}$ and $\{w_i^0\}$, having respectively N and M points, we define the grid in the common velocity interval $I = [\max(u_1^0, w_1^0), \min(u_N^0, w_M^0)]$. We then count the number of pixels encompassed within this interval for each of the two spectra, and we take the smallest of two numbers to be the cardinality of the common grid grid n_g. In this way we avoid oversampling in the rare cases where one spectra is observed with a smaller pixel density than the other. The spacing is then simply $|I|/n_g$, where $|I|$ is clearly the length of the interval. We finally rebin the transmitted fluxes onto the newly-defined pixel vector and we are set to compute the phase differences by standard Fourier analysis.

4.3 Calibrated Phase Analysis

In the previous section we described two ways of calculating phase differences from the observed Lyα forest in quasar pairs. The final goal of this calculation is to quantitatively assess the similarity of the measured phase distributions with those expected from the simulations for different IGM parameters, following the statistical formalism we devised in chapter 3.

However, we cannot calculate phase distributions directly from the simulated skewers. The simple reason is that the differences between the phase PDFs of data and simulations are also driven by non-astrophysical factors such as noise and resolution limit. If we want to exploit the sensitivity of phases to the filtering scale, we need first to understand and correct for the contribution of these disturbances. To this end, two different approaches are possible: find a way of subtracting them directly from data, or adding them to the simulations such that they are calibrated to the observations (*forward-modeling*). For the purposes of the Jeans scale measurement we choose to follow the latter, motivated by the simplicity of implementing forward-modeling in the context of our Bayesian machinery.

The calibration is done by creating, for each observed pair, an entire ensemble of simulated pairs with the same data properties, in particular the same transverse separation, the same noise amplitude and the same resolution. The next paragraphs of this section are dedicated to illustrate in details this procedure.

4.3.1 Transverse Separation

Two quasars separated in the sky by an observed angle ψ have a transverse distance dependent on their redshift. If we are studying Lyα absorption, the transverse separation between the coeval forest in the two spectra is an evolving function of the wavelength, since the sightlines are convergent toward us. The exact function depends on the cosmology, and can be written as

$$r_\perp(z_{\rm abs}) = D_A(z_{\rm abs)})\psi(1+z_{\rm abs}) \tag{4.14}$$

where $z_{\rm abs} = \lambda/\lambda_\alpha - 1$ is the Lyα absorption redshift and D_A is the correspondent angular distance. The variation of $r_\perp$ across our redshift bins is not negligible, especially for the longer chunks of forest, as figure 4.1 suggests. Since we know that phases are dependent on $r_\perp$, we should take this fact into account. Calculating the optical along arbitrary directions in the simulation would be complicated to implement, so we prefer to follow an alternative strategy. We keep choosing sightlines parallel to the box coordinates, but

for each observed pair we compute a full ensemble of synthetic pairs with separations uniformly distributed over the range covered by $r_\perp(z_{\rm abs})$ within the redshift limits of the chunk. In practice, if the coeval Lyαforest of the pair lies between $z_{\rm min}$ and $z_{\rm max}$, we simulate 400 pairs randomly located in the box and with separation $\{r = r_\perp(z_i)\}$, where the 400 redshifts z_i are logarithmically spaced between $z_{\rm min}$ and $z_{\rm max}$. The logarithmic spacing is chosen to achieve linear spacing in $v(z)$, which is the coordinate on which Fourier coefficients are calculated.

4.3.2 Resolution

We know that phases have the mathematical property of being invariant under convolution with symmetric kernels. For this reason one may think that no corrections for the resolution are required. Unfortunately, this invariance does not hold in presence of noise, analogously to a general deconvolution problem. In fact, phase scattering due to noise is enhanced at high-k where the signal from the forest is suppressed due to resolution limit. Phases lose their alignment and their intrinsic probability distributions is flattened depending on the noise level and the resolution kernel. A more precise description of the relation between noise, mode amplitudes and phase dispersion is given in § 4.4. The conclusion we draw from this argument is that the *combined* effect of resolution and noise must be always taken into account, unless the data have exquisite signal-to-noise ratio. We also conclude that when the power of the signal drops due to resolution, our measurement are unreliable or in the best case useless, because we are only sensitive to noise. For this reason we must set as an upper limit on the usable $k-$range near the resolution cutoff, which is $k_R = 1/\sigma_R \approx 2.355/\mathrm{FWHM}$, where σ_R is the width of the resolution kernel and FWHM the relative width at half maximum. The exact choice of this threshold will be discussed in the next chapter (§ 5.1.5).

In the forward-modeling approach we convolve the simulated spectra with a Gaussian, with the FWHM defined by the resolution of the spectrograph. Although the resolution is wavelength-dependent, we use a constant width for each Lyαforest chunk which is specified in table 4.1. This width corresponds to the FWHM at the average wavelength of each chunk, where the average is defined as the median in velocity space, which can be shown to be

$$\bar{\lambda} = \frac{\lambda_1 \lambda_2 \ln(\lambda_2/\lambda_1)}{\lambda_2 - \lambda_1}, \tag{4.15}$$

where λ_1 and λ_2 are respectively the minimum and the maximum observed wavelength of the chunk.

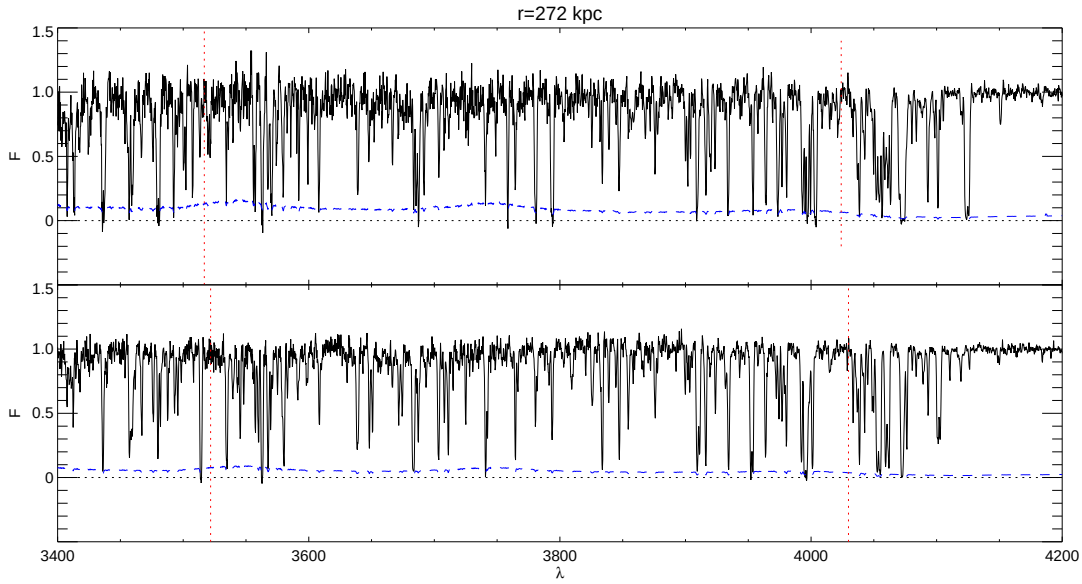

FIGURE 4.2: Transmitted flux as a function of wavelength (in Å) for a pair in our sample. The two spectra have been observed with MAGE, and the two quasars are both at $z = 2.38$ and have a comoving separation of 272 kpc. The noise level is marked with a blue dashed line. Both the transmitted flux and the noise have been renormalized by the continuum emission. The red dotted lines delimit the wavelengths of the Lyα forest in the two spectra.

4.3.3 Noise

The quality of the data is significantly variable within our sample, with the S/N per Agström varying in the range between approximately 5 and 35 (we define the signal-to-noise of a chunk as the minimum between those of the two companions, calculated in the redshift interval of the used Lyα forest). As discussed above, noise alters phases by blurring the alignment and by flattening their distributions. For a fixed S/N, this is analogous to convolving the intrinsic phase probability function at a given k and $r_\perp$ with a kernel determined by the noise power (see § 4.4 for more details). Phases calculated from pairs with different S/N are scattered at different levels and so their distributions are not directly comparable, demanding a specific calibration for each object.

Another complication stems from the wavelength-dependent nature of the noise, which is typically higher at smaller λ (see for example figure 4.2). Although the variation is not strong and this is probably a second order effect, we model it by applying to the synthetic spectra the same noise vector estimated, pixel by pixel, in real data. This operation is complicated by the fact that skewers drawn from the simulation all have the same length (50 Mpc/h) and the same pixel spacing, while each observed spectrum has its own. We solve this problem by periodically replicating each simulated spectrum until its size matches that of the forest chunk on which it is calibrated (see figures 4.3 and 4.4).

The procedure of extending the segments of Lyα-forest reduces the fundamental frequency in Fourier space and thus increases the density of modes. This new modes would introduce spurious and redundant information which have a substantial effect on phase distributions. The correct way of calculating phases after the extension is to split the final spectra in chunks of size equal or smaller than the box length (the number of replications may be a non-integer number), and extract their phases separately. The sample created in this way can then be used to determine the probability distribution function. Although there is still redundant information, such redundancy will only affect the variance and not the mean of phase probability, since this procedure is equivalent of resampling the same region of space more than once. The same chunking technique cannot be done on data, since phases from consecutive segments of Lyα forest are likely to be correlated, and such correlation is not included in our likelihood estimator. Therefore phases are extracted from the observed spectra by applying the Fourier analysis to the full length of the chunks.

The flux of the spectrum obtained after the periodical extension is finally rebinned into the same pixel grid of the observed spectrum, which is always coarser than the one used in our simulation. Once this is done, we are set to generate Gaussian noise matched pixel by pixel to the estimated wavelength-dependent noise of data. The noise in the Lyα forest is always renormalized by the continuum.

4.3.4 Forward-Modeling of the Simulation

The steps illustrated in the three previous paragraphs constitutes the *forward-modeling* of our simulation. This consists in applying to the simulated transmitted flux the same alterations that affect the real spectra when they are observed through a telescope with finite resolution and integration time. Forward-modeled simulations can be safely compared to observations, and allows to implement the same Bayesian formalism described in chapter 3. The forward-modeling need to be tailored separately for each pair, and must be applied to all the IGM models that we want to test. It is useful to summarize the general procedure that we follow to perform the phase-difference analysis on our data sample.

Suppose that we want to calibrate a model T_0, γ, λ_J to estimate the PDF of phases measured from the Lyα forest of two quasars Q_1 and Q_2 separated in the sky by an angle ψ, in the a redshift bin $Z = [z_1, z_2]$. This operation can be structured as follows:

- we determine the overlapping portion of the Lyα forest of the two QSOs which intersects Z. This segment will have a comoving separation varying with redshift

as $r_\perp(z) = D_A(z)\psi(1+z)$ (see section § 4.3.1). From now on we will use the velocity-space notation $r_\perp(v)$ for the impact parameter and $F_1(v), F_2(v)$ for the transmitted fluxes of the two spectra.

- we generate 400 pairs from the simulated box distributed in transverse separations $r_\perp$ depending on $r_\perp(v)$ as described in § 4.3.1.
- As in chapter 2, the optical depth is globally renormalized in order to match the observed mean flux.
- All the 400 pairs are forward-modeled according to the properties of Q_1 and Q_2. In each simulated pair one companion is associated to Q_1 and the other to Q_2, which have in general different resolutions and S/N. This is done through the following four steps:

  1. convolving the flux skewers with a gaussian kernel, with FWHM defined by the spectral resolution.
  2. Replicating them periodically until they equal the length of the observed segments of forest.
  3. rebinning the simulated flux into the same pixel grid of data.
  4. adding gaussian, uncorrelated noise to the simulated flux. The S/N matches at each pixel the one estimated in the observed spectrum.

- We finally calculate phases from the skewers and estimate the wrapped-Cauchy concentration parameters ζ at each bin in k. We predict in this way the probability distribution of phases $P(\theta; k)$ as a function of k for the considered pair. Note that having rebinned the skewers at step 3, we have to calculate phases with the LSSA method as we do with data.
- Following the method elaborated in § 3.1.6 we write the likelihood as

$$\mathscr{L}(\theta|M) = \prod_i P_{\rm WC}(\theta(k_i)|\zeta(k|M)) \tag{4.16}$$

  where $\theta(k_i)$ are the phases of the real pair and $\zeta(k|M)$ the wrapped-Cauchy parameters estimated for the model M via forward modeling. Note that differently than equation 3.12 there is no index over $r_\perp$, because this likelihood refers to only one observed pair. This likelihood is evaluated only below the limiting wavenumber $k_{\rm MAX}$, which is the minimum between k_R (resolution limit) and 0.1 s/km (limit set by metal contamination).

This procedure is then repeated for each observed pair, for each IGM model and at each redshift bin. This provides values of the likelihood function for all the pairs through the

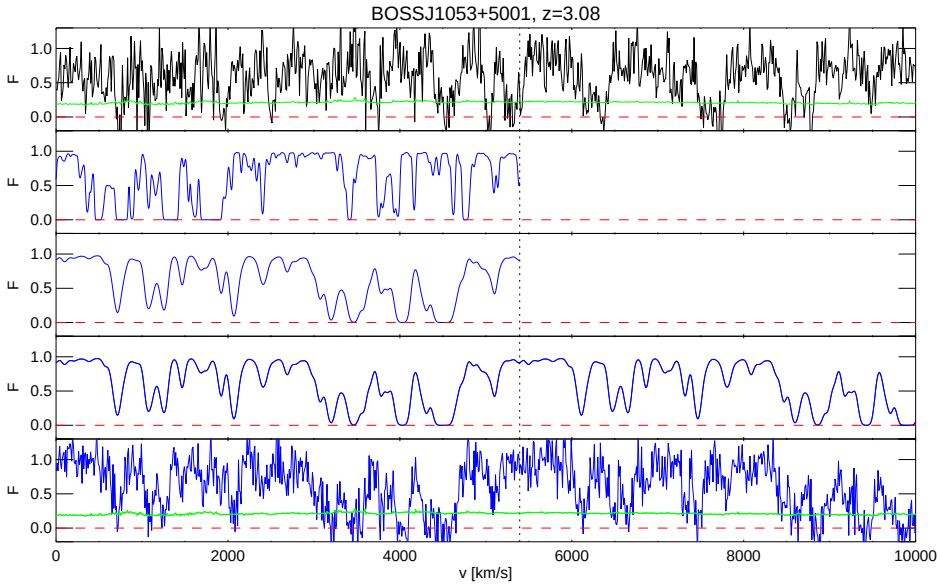

FIGURE 4.3: Example of our forward-modeling procedure at z=2.4. The top panel shows an ESI spectrum with resolution of 64 km/s at FWHM and average signal-to-noise ratio of 7.7 per Å. A random sightline from the snapshot at $z = 3$ is selected and plotted in the second panel. This synthetic spectrum is then smoothed according to the data resolution (third panel), extended periodically to match the length of the observed segment and rebinned into the same pixel grid of data (fourth panel). Finally, Gaussian noise is added according to the estimated error at each pixel (green line), which completes the process of forward-modeling (bottom panel). The vertical dotted line marks the length of the simulated box. Phases are calculated in the simulation for each of the segments separated by the vertical lines, while in data they are extracted from the full chunk.

whole parameter space, allowing us to make inference on the thermal properties at the different redshifts.

4.4 Effect of Noise on Phase Distribution

In this section we give an analytic expression for the scattering of phases in the presence of noise. It is not used in our forward-modeling scheme, where the PDFs are calculated after adding noise to the skewers, but it provides useful insights on the behavior of phase differences on real data and it will be useful for further discussion in the next chapter.

We assume that the noise is described by Gaussian random fluctuations with a constant power spectrum $P_N(k) = P_N$. Let us consider a mode of a noiseless spectrum at wavenumber k with amplitude $\rho(k)$ and, without loss of generality, phase $\phi_0 = 0$. Noise can be modeled as a stochastic variable $z_N = a_N + ib_N$ in the complex plane, with a uniform distribution in phase and a Gaussian distribution in modulus. It can be shown that the probability function of the phase ϕ of the complex stochastic variable

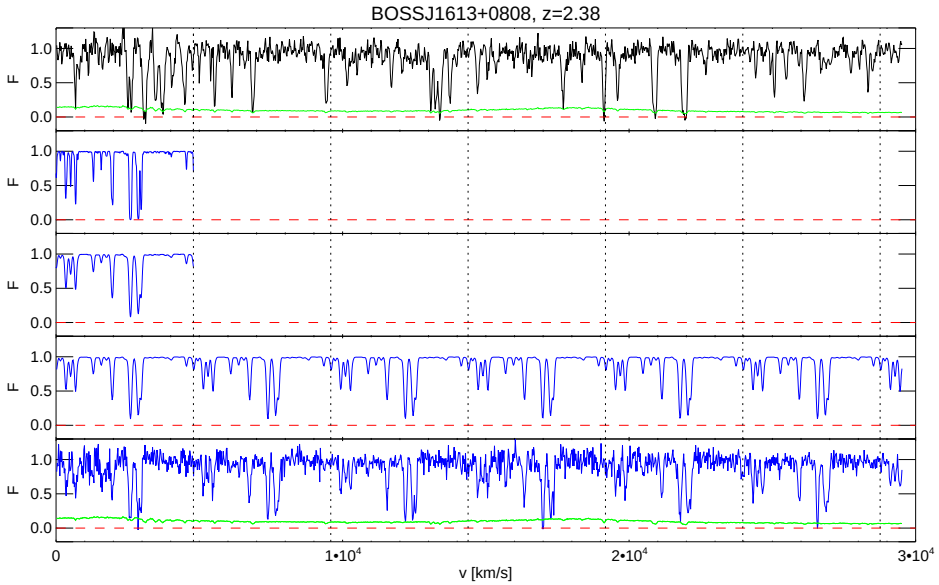

FIGURE 4.4: Same as Figure 4.3 but at redshift $z = 2$. The top spectra was observed with MAGE at resolution of 62 km/s (FWHM) and signal to noise of 17.7 per Å. Compared to the previous plot the data quality is higher, there is sensibly less absorption and the forest segment is longer, so we need to replicate the skewer more times.

$F = F_0 + F_N$ is

$$p_N(\phi|\eta) = \frac{e^{-1/2\eta^2}}{2\pi} + \frac{\cos\phi}{\sqrt{8\pi\eta^2}} e^{-\frac{\sin^2(\phi)}{2\eta^2}} \operatorname{erfc}\left(-\frac{\cos\phi}{\sqrt{2\eta^2}}\right), \tag{4.17}$$

where we define the "noise parameter" $\eta = \sqrt{P_N}/\rho(k)$. The full derivation is given in appendix , but it is interesting to note that the distribution follows the expected behavior in the limiting cases. When the noise is very high ($\eta \to +\infty$), it reduces to a flat distributions $p_N(\phi) = 1/2\pi$, meaning that phases have totally lost the original coherence information. Conversely, when the signal dominates ($\eta \to 0$), p_N is well approximated by a Gaussian in $\sin\phi$ with variance η^2.

Formula 4.17 expresses the noise scattering of the phase of a single mode. If we want the dispersion of the phase difference of two homologous modes in a pair, we would need to calculate the distribution of the sum of the noise phases ϕ_1 and ϕ_2, and that is given by the convolution of the p_N for the two modes. We do not attempt to derive a general expression for this convolution, but we note that in the limit of low noise ($\eta \ll 1$) it reduces to a convolution of two Gaussians, which leads to another Gaussian with the variances added in quadrature $\eta^2_{\Delta\phi} = \eta^2_{\phi_1} + \eta^2_{\phi_2}$.

Chapter 5

Results

We have now developed all the tools that are needed to attempt a measurement of the Jeans scale on quasar pairs.

In chapter 3 we showed that phase difference analysis applied to close quasar pairs is capable of optimizing the sensitivity to the spatial coherence of the low-density IGM and minimize the degeneracies with the thermal parameters T_0 and γ. We devised a statistical procedure that allows a rigorous probabilistic inference in parameter space, estimating a potential precision of $\lesssim 5\%$ on the filtering scale λ_J with a realistic sample of 20 full high-quality pairs.

The current data set of quasar pairs described in the previous chapter does not reach the same level of quality and quantity, but it allows the first step in the direction of a high-precision measurement of the Jeans scale, setting for the first time constraints on λ_J. We illustrated how we calibrate the simulations to take into account the wide range of noise and resolution of the spectra we want to analyze, following a forward-modeling approach § 4.3. Such calibration enables the prediction of the expected phase difference distributions for each of the pairs in our sample, given a theoretical (i.e. noiseless) model for the IGM.

The combination of forward modeling and statistical analysis of phase-differences lead to the results that we present in this chapter. The details of the implementation of the statistical analysis are specified in § 5.1, and the constraints on the parameters at the three redshift intervals are given in § 5.2. Finally we test the robustness of our results for a series of possible bias source (§ 5.3).

5.1 Implementation of the Statistical Analysis

5.1.1 Simulation

This measurement was conducted using a Nbody simulation analogous to that described in § 2.1.2, but on a smaller box, capable of resolving smaller Jeans scales. This choice is motivated by preliminary results that indicated unexpectedly low values for λ_J. We use a cube of 30 Mpc/h of size and 2048^3 DM particles, which according to the criterion established in appendix A enables to study Jeans scales of $\lambda_J \approx 15$ kpc. We also updated the cosmological parameter to Planck results [Ade et al., 2014], i.e. $\Omega_\Lambda = 0.68, \Omega_m = 0.32, h = 0.67$. We focus on three snapshot at $z = 2, 2.4, 3$, approximately at the centers of the redshift intervals in which we bin the data.

5.1.2 Parameter Grid

As explained in chapter 2, the interpolation of statistics in parameter space (i.e. the *emulator*) takes great benefit from a careful designing of the training grid. In particular it is important that the grid has good *filling* properties in parameter space, meaning that all subspaces are sampled homogeneously and as densely as possible, which is not achieved neither with regular nor with randomly generated sets of points. To conduct our data analysis we employ e parameter grid of $n_m = 405$ points in the $T_0 - \gamma - \lambda_J$ space. The fact that we use a smaller grid than in our theoretical study is motivated by the fact that we expected a smaller precision from data compared to the perfect spectra of simulation. It is anyway possible to refine such grid if required, either *a posteriori* because the results do not converge or because of an improvement of the data sample.

We use a more efficient algorithm to generate a space-filling grid compared to what we employed in Chapter 3. The properties of the parameter grids are the following:

- we use n_m different values of T_0 and n_m different values of γ uniformly distributed within the chosen range.We have only $n_J = 45$ values for the Jeans scale λ_J, since it is more computationally expensive to vary than the other two parameters. Each of these value is used 9 times.
- we divide the T_0 and γ ranges in 9 regular intervals (which we will call *segments*), each with 45 points, and the λ_J range in 5 segments, each with 81 points. This subdivision defines in the $T_0 - \gamma$ 2d subspace a grid of 9×9 cells (from now on *quadrants*). Each quadrant of this plane is populated with 5 points. Analogously,

we define two 5×9 grids in the λ_J-T_0 and in the $\lambda_J-\gamma$ planes, with each quadrant having 9 points.

- the full 3d space $T_0-\gamma-\lambda_J$ is now naturally divided in $9\times9\times5$ cells. Each of them contains one point.

The three points express respectively the conditions for 1d,2d and 3d homogeneity for all possible subspaces. A precise description of the details on how this grid is devised are beyond the scope of this work.

5.1.3 Likelihood

In chapter 3 we have introduced the likelihood estimator for the phase difference statistic as

$$\mathscr{L}(\{\theta\}|M)=\prod_{k,r_\perp}P_{\rm WC}(\theta(k,r_\perp)|\zeta(k,r_\perp|M)) \tag{5.1}$$

where $\theta(k,r_\perp)$ is the phase difference between the k-modes of a quasar pair with impact parameter $r\perp$, and ζ is the concentration parameter of the wrapped-Cauchy distribution. ζ depends on $k,r_\perp$ and on the IGM model $M=\{T_0,\gamma,\lambda_J\}$. We now apply this likelihood function to the phases calculated from real pairs with the least-square spectral analysis technique (see § 4.2). The ζ parameters are obtained through our grid of DM-based models, after careful calibration of noise and resolution (as described in § 4.3). The range in k available for the analysis depends on one hand on the size of the simulated box, which sets the lower mode, and on the other hand on the resolution of the instruments, that cuts the signal of high-k modes. Regardless of the resolution, we never consider wavenumber greater than $k=0.1$ s/km due to metal contamination. A detailed description on how we establish the limiting k as a function of the resolution will be given in § 5.1.5. The ζ parameters are calculated for each $r_\perp$ and for 13 different bins in k-space. This bins are equally spaced in $\log k$ between $k=0.005$ s/km and $k=0.1$ s/km.

5.1.4 Interpolation

Analogous to our study in chapters 2-3 we use the models in the discrete set of 405 points in the grid to make prediction in the continuum of parameter space, by means of Gaussian-processes interpolation. Differently than what we have done there, we do not use the emulator to interpolate the *full* statistic that we are using, i.e. phase difference distributions. We follow the simpler method of calculating the likelihood at each point of the parameter grid with equation 5.1 and we interpolate only that single number. This choice allows a much faster calculation than emulating the full statistic, which is

not required since we only need the likelihood probability to perform the measurement. Moreover, if the results are converged they should be independent from the choice of the interpolation algorithm.

The convergence of the emulator is achieved when the density of the training grid in parameter space is consistent with the smoothness of the interpolated variable. If this is not fulfilled, a refinement of the parameter grid is necessary. In the case of the likelihood, the smoothness depends on the dimension of the sample: the more constraining the data, the higher the requirements on the grid density. To make a simple example, if the inferred 1σ region contains only one point of the parameter grid, the size and the shape of our confidence levels would entirely depend on the choice of the interpolation parameters, and could not be trusted. In the context of Gaussian processes, the error would be set by the correlation length in parameter space which determines the correlation matrix. It is therefore required that we populate with enough grid points the regions of parameter space where the posterior probability is not negligible. Instead of determining a general criteria of "filling density", we adopt a simple *a posteriori* consistency check: we use correlation lengths larger than the typical model-model separation in parameter space and check that the results of the measurement do not sensibly change when we vary these parameters (§ 5.3.4.2). In typical Gaussian process implementations this hyperparameters are determined by maximizing the likelihood of the sampled points, i.e. in the training grid. We consider this approach too arbitrary for our study, since we use GP for interpolation purposes but we do not have any good reason to believe that the values of the likelihood are effectively drawn from a Gaussian process.

5.1.5 Resolution Limit on $k_{||}$

Based on previous study on the line-of-sight power spectrum [McDonald et al., 2000], we argued in chapter 3 that the Fourier modes in the forest with wavenumber $k_{||} > 0.1$ s/km should be excluded from the analysis because of contamination of narrow metal lines. We also know that according to the instrument resolution the power of the signal drops exponentially as $\exp[-(k\sigma_r)^2]$, assuming that the resolution can be modeled as a constant Gaussian kernel with width σ_r ≈FWHM/2.355. As the noise has a white power spectrum, it will be the dominating source of power beyond this cutoff, erasing the sensitivity to the Jeans scale. If we trusted completely our forward-modeling procedure, these noisy high-k modes should not represent a problem, since they are consistently calibrated in the simulation. In that case, including them in the analysis would only add uninformative phases with flat distributions, with no effect on the final inference. However, it could be that our assumption on the shape of the resolution kernel (Gaussian with constant width) or on the properties of the noise (white Gaussian noise) are not

precise enough in the delicate conditions where phase distributions are more sensitive to these effects than to the thermal parameters because of a low signal. Therefore we find preferable to remove these modes from the likelihood,since they could lead to biased results if our forward modeling is imprecise.

Our goal is then to define a "safe" dynamic range $k < k_{\rm res}$ to which applying phase analysis. The definition of $k_{\rm res}$ should take into account the S/N level: we have shown in chapter 3 that phases are invariant under convolutions with symmetric kernels, meaning that the effect of resolution should be irrelevant in the limit of S/N$= \infty$. It is reasonable to believe that higher quality data should allow a more extended dynamic range than noisier spectra. This argument can be better motivated using formula 4.17. The parameter that regulates phase noise as a function of k, and thus the alteration of phase PDFs, is the "noise parameter" η defined by

$$\eta(k)^2 = \frac{P_N(k)}{\rho^2(k)}, \tag{5.2}$$

where $P_N(k)$ is the noise power spectrum and $\rho(k)$ the *amplitude* of the examined Fourier k-mode. Regardless of the resolution, η is always zero in absence of noise, but in the realistic case of finite noise its value increases exponentially with k. In fact, we can assume that the noise follows a white power spectrum $P_N(k) \propto ({\rm S/N})^{-2}$, and that the squared amplitudes of the modes are set (on average) by the LOS power spectrum and by the resolution as

$$\langle \rho^2(k) \rangle = P_{\rm LOS}(k) \exp[-(k\sigma_r)^2]. \tag{5.3}$$

With these assumptions we can write the noise parameter as

$$\eta(k)^2 \propto P_{\rm LOS} \frac{\exp[(k\sigma_r)^2]}{({\rm S/N})^2}, \tag{5.4}$$

which diverges exponentially at high k, as claimed above. $P_{\rm LOS}$ introduce a second cutoff due to thermal broadening and Jeans smoothing.

The last equation suggests a criterion to fix the maximum k as a function of the signal-to-noise ratio. Since the quality of phase PDFs is set by η, we can request this parameter to be smaller than a certain threshold $\bar{\eta}$. By imposing the condition $\eta > \bar{\eta}$ and by inverting equation 5.4 we obtain

$$k < k_{\rm res} \equiv \frac{1}{\sigma_r} \sqrt{2 \log({\rm S/N}) - \xi}, \tag{5.5}$$

where the free parameter ξ encompasses all the proportionality factors, the LOS power and the choice η_0. For simplicity, we are also assuming that the intrinsic power of the forest is flat, which is reasonable as long as the thermal cutoff occurs at higher k than the typical resolution cutoff of our sample. To take into account the noises of both spectra

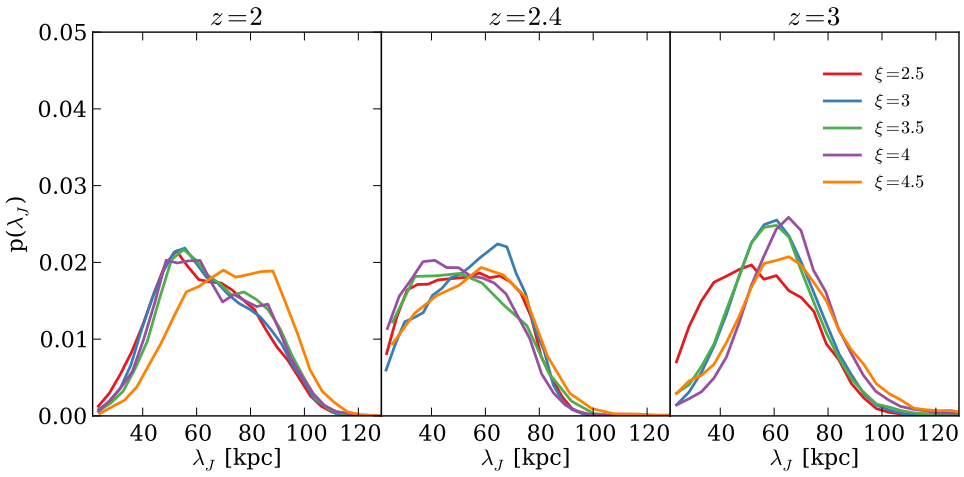

FIGURE 5.1: Calibration of ξ using the λ_J posterior distributions. We run MCMCs assuming a set of values of ξ for the three redshift bins in our analysis. This set ranges between 2.5 amd 4.5, as the legend shows. We select the value of ξ according to the convergence and the width of the posterior distributions: high values of ξ are more conservative and might reject constraining modes, while a low ξ would include uninformative distributions or give rise to bias due to wrong noise modeling and degrade the constraints. An example of such a degrade can be seen in the right panel when adopting $\xi = 2.5$ (red line). Based on this plot we opt for $\xi = 4$, for which the posteriors are reasonably converged in all the three redshift bins.

of a pairs, we can assume that η_1 and η_2 of the two companions add in quadrature, i.e. $\eta^2 = \eta_1^2 + \eta_2^2$. This is justified by the fact that phase dispersion is well approximated by a Gaussian when η is small (see § 4.4) and that the dispersion of phase differences is given by the convolution of the dispersion kernels of the two individual phases. In this case, based on equation 5.4 we can define the effective signal-to-noise of a pair as

$$\mathrm{S/N} = \frac{(\mathrm{S/N})_1(\mathrm{S/N})_2}{\sqrt{(\mathrm{S/N})_1^2 + (\mathrm{S/N})_2^2}} \tag{5.6}$$

where $(\mathrm{S/N})_1$ and $(\mathrm{S/N})_2$ are the signal-to-noise ratios of the two companions.

We leave for future work a better treatment and optimization of $k_{\mathrm{res}}(\mathrm{S/N})$ at different redshifts. Note that our criterion authomatically sets a lower cut on the signal level, by demanding a positive argument of the square root:

$$\mathrm{S/N} > \exp(\xi/2). \tag{5.7}$$

Higher values of ξ are more conservative, since they fix the S/N cut at higher levels and they are more restrictive with respect to the k dynamic range.

We determine ξ by looking at the posterior distributions of λ_J obtained from a series of MCMC runs that assume different values of ξ (figure 5.1). If our forward modeling

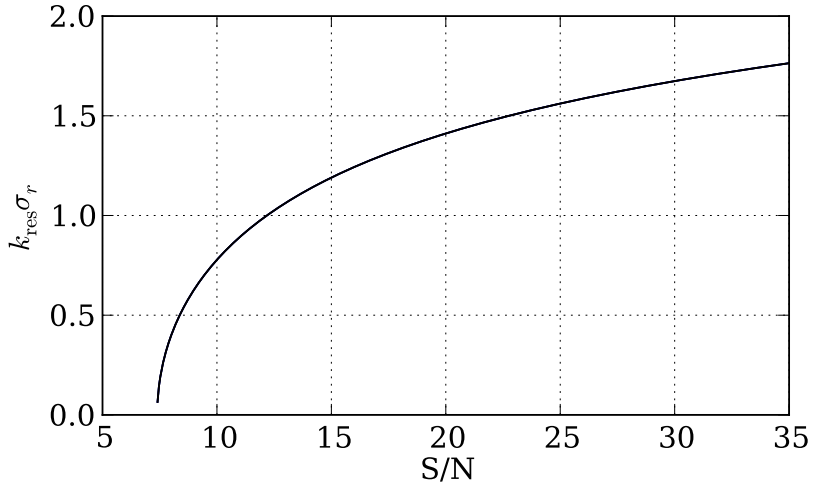

FIGURE 5.2: Dependence of $k_{\rm res}$ on the S/N of a pair for the adopted value of $\xi = 4$. The lower cut on S/N is set to $\exp(\xi/2) \approx 7.4$.

is correct, as we decrease ξ we include more modes in the likelihood and therefore we expect one of the following behavior:

- The new modes retain information about the Jeans scale and thus our accuracy improves;
- The new modes are dominated by noise and our constraints do not change.

If, on the other hand, the precision degrades at low ξ (i.e. the width of the posterior distribution increases), it is likely that our modeling of the noise and resolution is not fails at the noisiest modes. Based on this argument and on the results shown in figure 5.1 we adopt the parameter $\xi = 4$, which is the most conservative value at which the posteriors look reasonably converged in all the redshift bins. Figure 5.2 explicitly shows the dependence of the maximum wavenumber $k_{\rm res}$ as a function of S/N when we set $\xi = 4$. The threshold on the signal to noise is $(S/N)_{\rm min} = \exp(\xi/2) \approx 7.4$.

5.2 Results

5.2.1 Phase Distributions of Real Pairs

It is useful to directly look at the phase difference distributions of quasar pairs, since on them is based our inference on the Jeans scale. The main problem in doing that is acquiring an homogeneous and statistically significant sample of phases in order to construct the probability density function. Ideally, a PDF is meaningful if it is derived

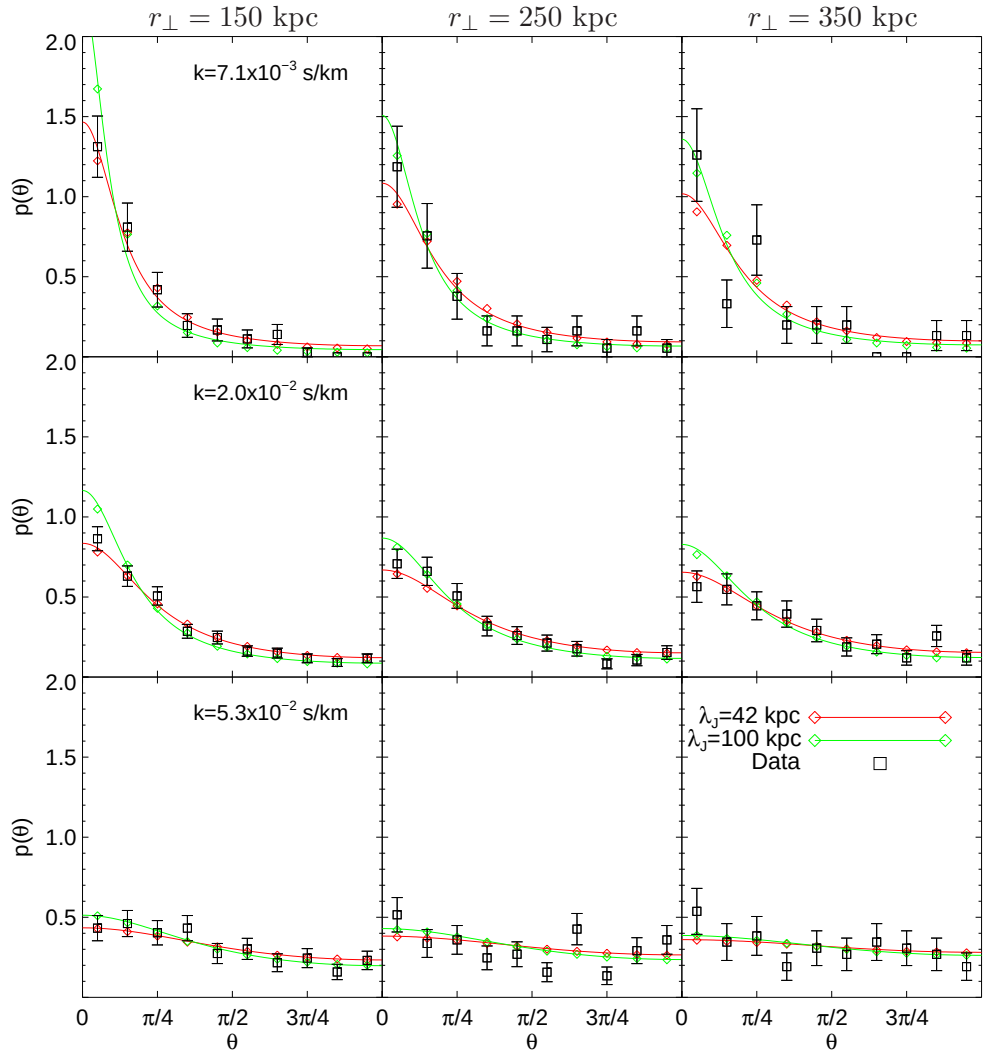

FIGURE 5.3: Phase difference probability distributions calculated from our data sample in the redshift interval $z \in [1.8, 2.2]$ (Black squares). In order to have a significant sampling we need to group phases into $r_\perp$ and $k_{||}$ bins. The panels from left to the right correspond to the $r_\perp-$intervals $[100-200]$, $[200-300]$ and $[300-400]$ kpc, respectively, while from top to bottom phases are split in the $k_{||}-$intervals $[0.005, 0.01]$, $[0.01, 0.04]$ and $[0.04, 0.07]$. The values marked in the plot refers to the central values of the bins (logarithmical in the case of $k_{||}$). It is worth reminding that the phases present in each bin are not homogeneous, i.e. they derive from data with different resolution and signal-to-noise ratio. We compare these distributions with the prediction of two fully forward-modeled simulations, with $\lambda_j = 30$ kpc (red diamonds) and $\lambda_J = 70$ kpc (green diamonds). Using the calibrated models instead of the original ones assures that the PDFs can be directly compared, since they are obtained from analogous pair samples.The solid lines are the wrapped-Cauchy fit of the models, which trace almost perfectly the underlying distributions.

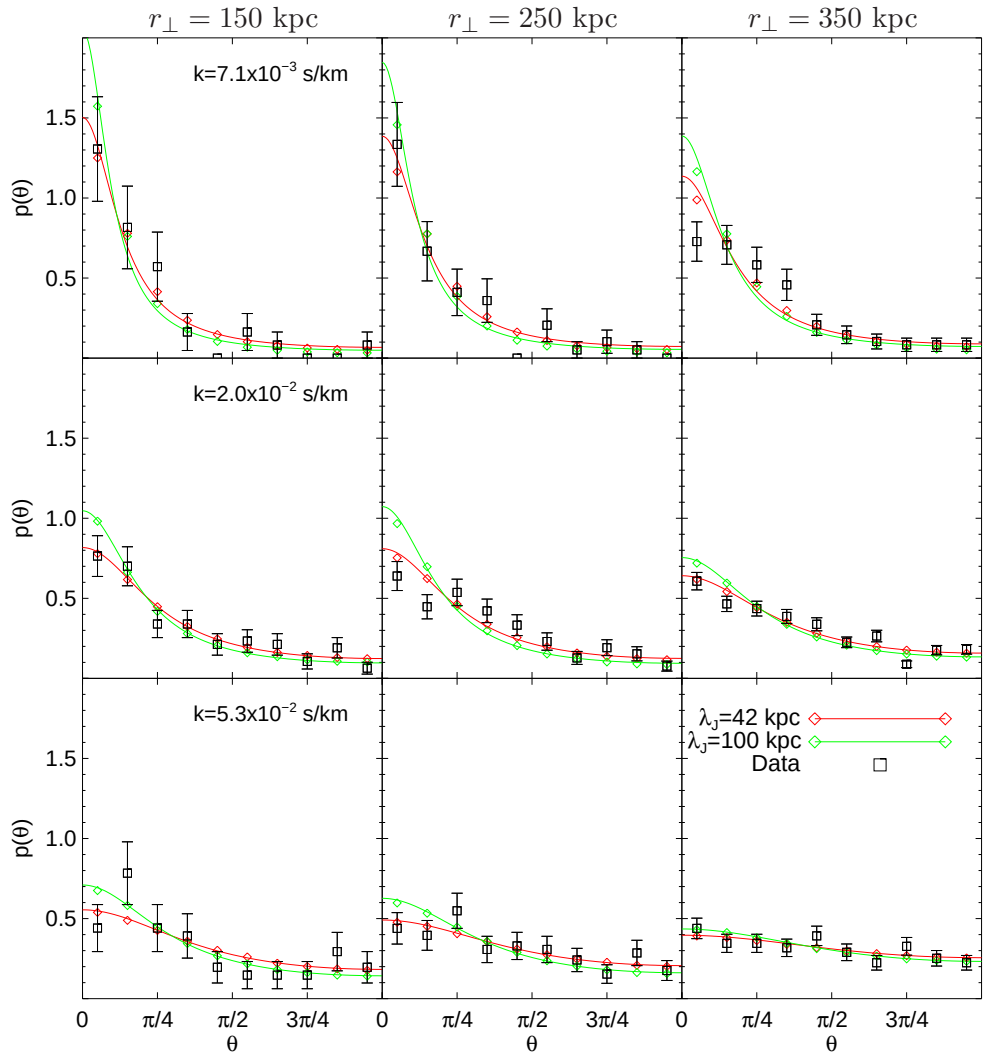

FIGURE 5.4: Same as figure 5.3 but for the redshift interval $[2.2, 2.7]$

from phases with the same physical and instrumental parameter, which in our case are $z, k_{||}, r_\perp$, the resolution and the signal-to-noise ratio. Unfortunately, our sample is not large enough to allow such an high-dimensional splitting, so we adopt a different approach. For each of the three analyzed redshifts we divide our data in relatively large $r_\perp$ bins ($\Delta r_\perp = 100$ kpc) and we calculate the phase distributions in three $k_{||}$ intervals $[0.005, 0.01], [0.01, 0.04]$ and $[0.04, 0.07]$ km^{-1} s. The bins in $k_{||}$ are defined somehow arbitrarily in order to have a comparable number of modes in each of them. The subsamples defined in this way are still hybrid, because they contain information obtain from data of sparse quality. However, our forward-modeling procedure allows us to produce analogous samples from the synthetic pairs of our simulations, which we can compare with observations. The results are shown in figures 5.3, 5.4 and 5.5.

The black squares are the PDFs obtained from data, while the red and the green diamonds are calculated respectively from the models $\{T_0 = 8000$ K$, \gamma = 1.3, \lambda_J = 42$ kpc$\}$

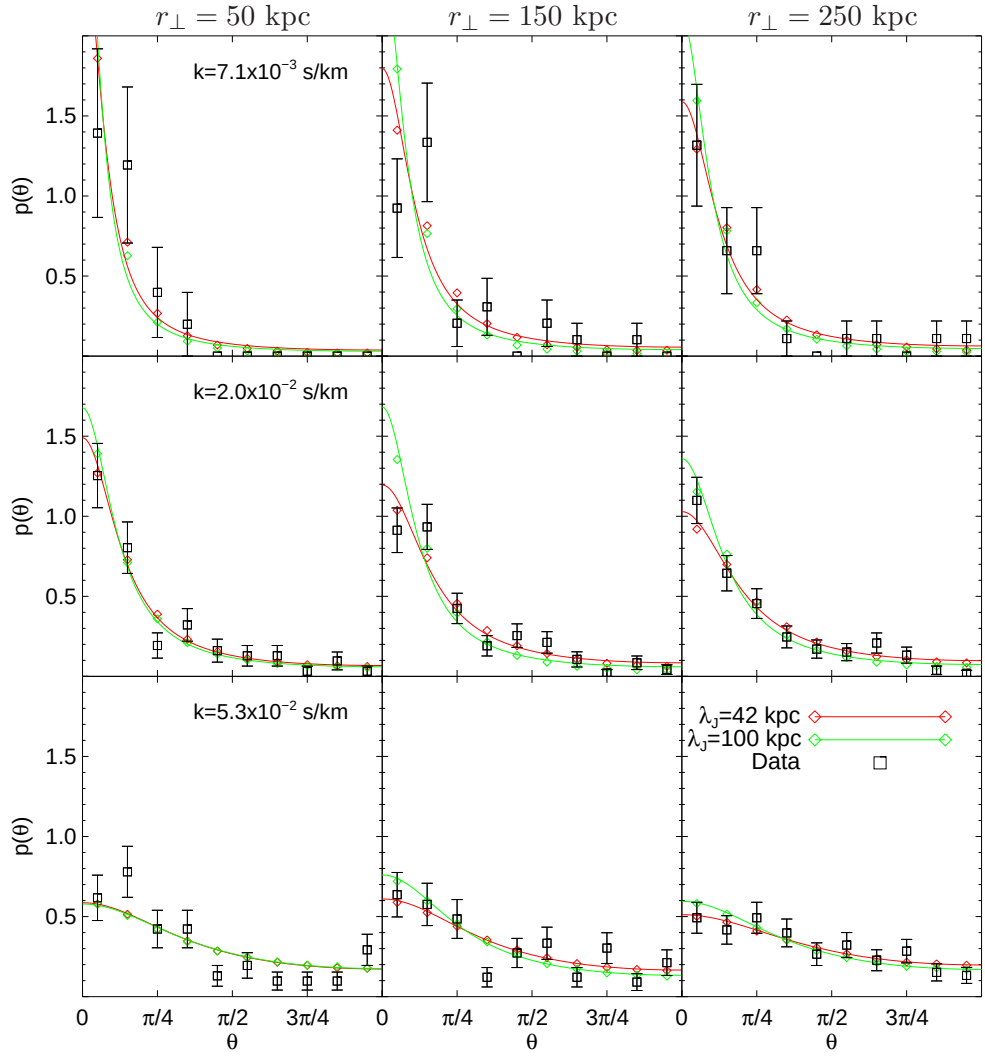

FIGURE 5.5: Same as figure 5.3 but for the redshift interval $[2.7, 3.3]$. Given the different distribution in $r_\perp$ we choose to plot phases in the bins $[0-100], [100-200]$ and $[200-300]$ kpc,from left to the right. Note that the sample is smaller at high redshift, so some bins are empty as in the top panels.

(which has the smallest Jeans scale of our grid) and $\{T_0 = 8000\ \mathrm{K}, \gamma = 1.5, \lambda_J = 100\ \mathrm{kpc}\}$. The pairs in the simulations are modified by adding noise and convolving with the resolution kernel so that they form a sample statistically comparable with the observed pairs. More precisely, for each pair we use 400 mocked pairs that are replicated proportionally to the length of the overlapping forest in data (see § 4.3 for details). This guarantees in particular that phases are correctly weighted accordingly to the extent of the observed segment of Lyα forest. The solid lines represent the best-fit of this mock phase distributions with the wrapped-Cauchy function, which falls overall very close to the actual values.

The behavior of phase PDFs follows the theoretical expectations described in chapter 3. Phases are generally more coherent at low $k_{||}$ and at small separations $r_\perp$. The shape of

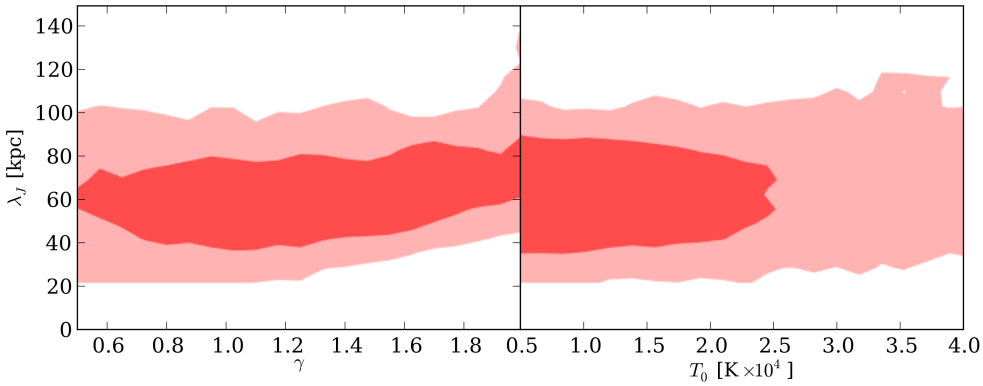

FIGURE 5.6: Constraints on the λ_J-γ and λ_J-T_0 planes at $z = 2$. The contours show the 65% and 96% confidence levels obtained by applying the phase difference statistic to our sample of quasar pairs between $z = 2.7$ and $z = 3.3$. As expected from our study described in chapter 3, there is no degeneracy neither with γ nor with T_0 at this redshift. Temperatures below 25000 K are slightly favored.

the measured distributions are also broadly consistent with a wrapped-Cauchy function, although the scatter looks still significant. We caution however that the errorbars plotted on data points are simple Poisson estimates and do not have a rigorous statistical meaning, given the hybrid nature of the sample inside each bin.

Despite of the illustrative purpose of these three figures, one can already see that the model with the smaller Jeans scale (red curve) provides in most cases a best fit to data points than the high-λ_J simulation. We may also guess that models with $\lambda_J < 42$ kpc, which would correspond to flatter distributions, are not ruled out by our dataset.

5.2.2 Constraints

We now present the results of the parametric study we have performed on data. The full Bayesian treatment allow us to draw quantitative conclusions from the phase-difference analysis.

The results in the redshift interval $z \in [2.7, 3.3]$ are shown in figure 5.6. The red contours are the confidence levels obtained from the MCMC run in the T_0-γ-λ_J space, projecting the posterior probability distribution in the λ_J-γ and λ_J-T_0 subspaces. These results meet our theoretical expectations in that the phase difference statistic at redshift 3 is insensitive on the temperature-density relationship. The confidence levels are horizontal as in figure 3.7, with which this plot can be compared. With the current sample we achieve a precision of about 30%. The full inference on λ_J, marginalized over the other parameter, is shown in figure 5.7. The estimated value of the Jeans scale is $\lambda_J = 66 \pm 20$ kpc, where the expected value and the error are calculated respectively as

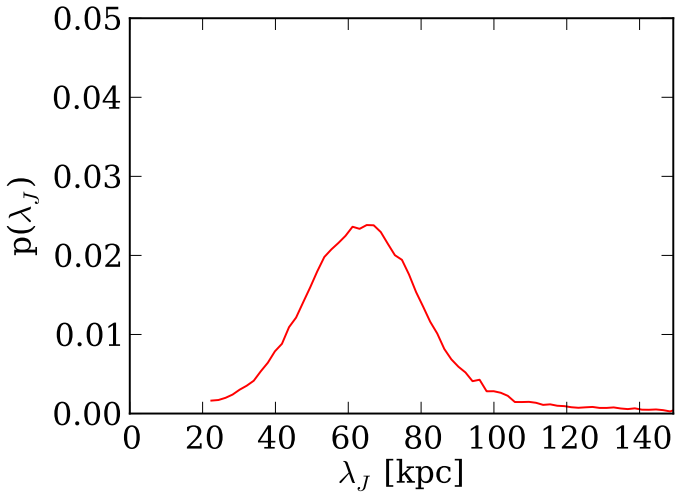

FIGURE 5.7: Accuracy on the Jeans scale measurement at $z = 3$. The plot shows the posterior probability distribution from the MCMC fully marginalized over the parameters γ and T_0. The expected value is $\lambda_J = 66$ kpc, and the estimated 1-σ error is $\Delta\lambda_J = 20$ kpc, giving a relative uncertainty of 30%.

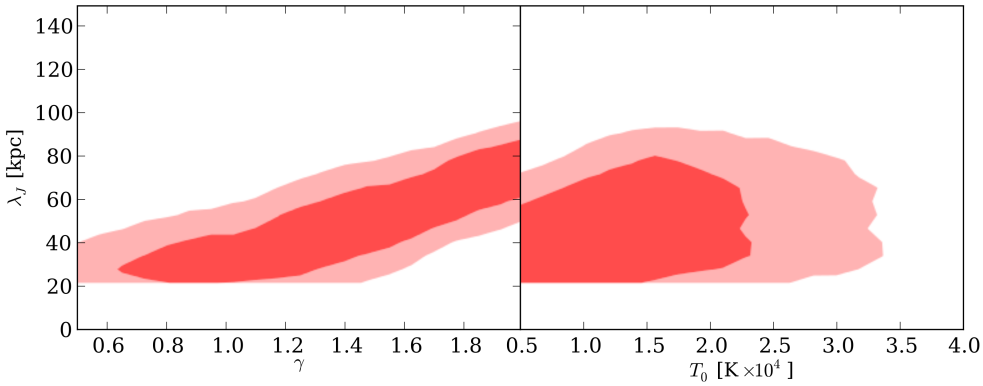

FIGURE 5.8: Same as figure 5.6, but in the redshift interval $[2.2, 2.7]$. Differently than $z = 3$, a tilt appears in the λ_J-γ contours, implying a degeneracy between the two parameters whose origin is still under research (see the text for a discussion). Similarly to $z = 3$, there is no degeneracy between λ_J and T_0, and low temperatures are favored. The sharp edge at $\lambda_J = 22$ kpc correspond to the lower border of our parameter grid.

the mean and the standard deviation of the MCMC chain. As a consequence of phase sensitivity, no constraints are set on γ and on T_0, except a very shallow preference for lower temperatures.

In figure 5.8 we present the same contours at redshift $z = 2.4$, obtained from the pair sample in the interval $z \in [2.2, 2.7]$. The most significant difference with $z = 3$ is the tilt of the confidence levels on the λ_J-γ plane, revealing a significant degeneracy between the two parameters. This degeneracy is an unexpected result, given our study at $z = 3$ and our understanding of phase difference statistic. It is somehow undesirable, since

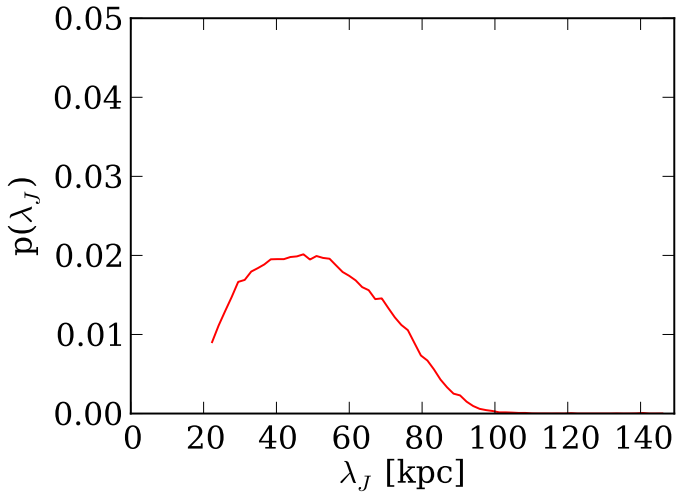

FIGURE 5.9: Same as figure 5.7, but at $z = 2.4$. As a consequence of the λ_J-γ degeneracy at this redshift the posterior is wider and not fully covered at the low-λ_J tail. The expected value and the standard deviations are $\lambda_J = 52$ kpc and $\Delta\lambda_J = 17$ kpc, respectively.

it loosen the constraints on λ_J such that at this redshift we cannot rule out extremely low Jeans scales ($\lambda_J < 20$ kpc !). We have not reached a clear understanding of how this degeneracy is originated. It could be that the forest at this redshift has different properties than at at $z = 3$, so the conclusion drawn in chapter 3 cannot be generalized. Alternatively, it may be that since the signal is smaller than at redshift 3, the phase statistic is more sensitive to noise and thus to the parameter η. According to equation 5.4, this would introduce a dependency on the LOS power spectrum, which in turn is sensitive to T_0 and γ. A precise explanation of this issue will require further quantitative analysis. However, we can notice that the degeneracy between λ_J and γ that we find with phase differences lies in a different direction than the degeneracy expected from the line-of-sight power spectrum(see figure 3.7). We can thus argue that crossing our results with line-of-sight measurement might significantly improve the constraining power of both statistics.

In the λ_J-T_0 plane the confidence levels are not significantly tilted, meaning that no degeneracy holds. Temperatures at mean density higher than 35000 K are excluded at 2-σ level.

The constraints on λ_J after marginalization of T_0 and γ are shown in figure 5.9. As a consequence of the degeneracy with γ, the posterior has a wider and flatter shape than its counterpart at redshift 3. The probability does not drop to zero at the low-λ_J tail, implying that filtering scales smaller than our current limit in the parameter grid ($\lambda_J \approx 22$ kpc) are not ruled out by the current measurement. Increasing the sample and understanding the λ_J-γ degeneracy are necessary steps in order improve the accuracy of

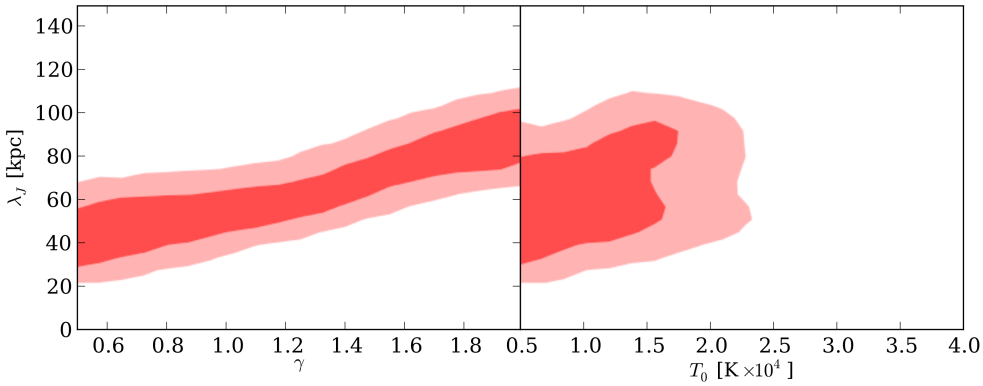

FIGURE 5.10: Same as figure 5.6 and 5.8, but in the redshift interval $[1.8, 2.2]$. The λ_J-γ degeneracy appears as at redshift 2.4, shifted towards higher values of λ_J. Temperatures at mean density above 25000 K are ruled out at 2-σ level.

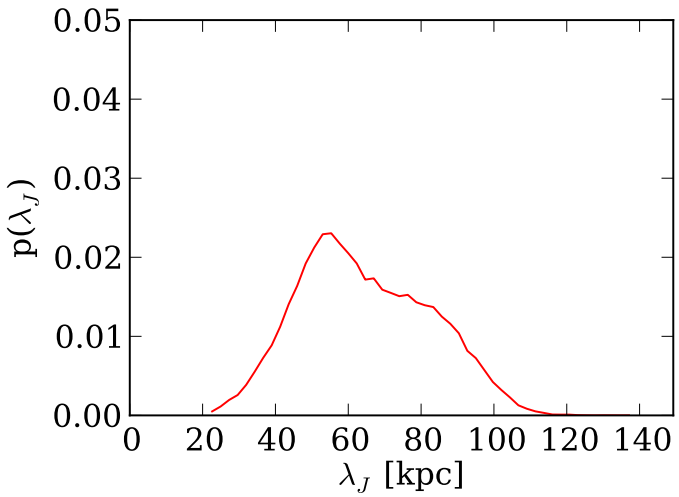

FIGURE 5.11: Same as figures 5.7 and 5.9, but at $z = 2$. The width of the distribution and the relative flatness of its top part are due to the degeneracy with γ. The expected value and the standard deviations are $\lambda_J = 64$ kpc and $\Delta\lambda_J = 17$ kpc, respectively.

the filtering scale estimation. The expected value and standard deviation are $\lambda_J = 52$ kpc and $\Delta\lambda_J = 17$ kpc, but we must stress that they are calculated within the parameter range covered by our simulation. Otherwise stated, we are assuming a prior $\lambda_J > 22$ kpc, which is not justified looking at figure 5.8.

The results at redshift 2 (figure 5.10) are qualitatively similar to $z = 2.4$. An analogous γ-λ_J degeneracy holds, but overall the contours lie at higher values of λ_J, excluding filtering scales smaller than 22 kpc at 2-σ level. The confidence levels are overall narrower than at $z = 4$, consistently with the larger sample size. Again, there is no degeneracy with T_0, and high values are more significantly ruled out ($T_0 < 25000$ K at 2-σ). The posterior probability distribution of λ_J is broadened by the degeneracy with γ, preventing a precise

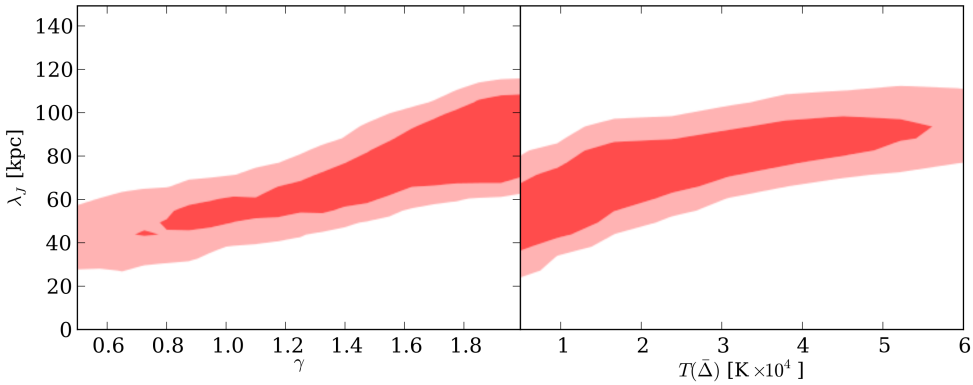

FIGURE 5.12: 65% and 96% confidence levels in the λ_J-γ and λ_J-$T(\bar{\Delta})$, where $T(\bar{\Delta})$ is the temperature at the "typical" overdensity of the Lyα forest Δ. In calculating these contours we assume a flat prior in $T(\bar{\Delta})$ instead that in T_0, which explains the difference of the left panels of this figure and figure 5.10 (see the text for a discussion). Differently than figure 5.10, the degeneracy with the temperature is now comparable with that with γ.

determination of the filtering scale. The expected value and standard deviation are $\lambda_J = 64$ kpc and $\Delta\lambda_J = 17$ kpc, implying a relative precision of 27%. The distribution however is not symmetric, and the highest probability is reached at $\lambda_J = 55$ kpc.

It might sound puzzling that the filtering scales is degenerate with the index γ of the temperature-density relationship, but not with the temperature at mean density T_0. As stated above, understanding this degeneracy will require further study, but we can argue that the Lyα forest at redshift 2 is not very sensitive to T_0 because it probes density significantly higher than the mean. It is therefore interesting checking whether a degeneracy holds after reparametrizing the T-ρ relationship with respect to the typical overdensities of the Lyα forest $\bar{\Delta}$ and its temperature $T(\bar{\Delta}) = T_0\bar{\Delta}^{\gamma-1}$, or simply $T_{\bar{\Delta}}$. $\bar{\Delta}$ is not precisely defined or measured, but an indicative value has been estimated in Becker et al. [2011] in the context of a measurement of the IGM temperature, giving $\bar{\Delta} \approx 4.11$ at $z = 2$. Figure 5.12 shows the confidence levels obtained after this transformation in parameter space. We find that the degeneracy of λ_J with $T(\bar{\Delta})$ is indeed as significant as the one with γ.

For consistency, in doing this study we adopt a flat prior in $T(\bar{\Delta})$ instead of a flat prior in T_0. This choice explain the slight difference of the left panel of figure 5.12 with its equivalent in figure 5.10. since the Jacobian of the parameter transformation is $\partial T_{\bar{\Delta}}/\partial T_0 = \bar{\Delta}^{\gamma-1}$, the probability transform according to $p(T_0) = p(T_{\bar{\Delta}})$. This last relation implies that a flat prior in $T_{\bar{\Delta}}$ implies higher probabilities at high γ compared to the flat prior in T_0 that was assumed in figure 5.10.

5.3 Consistency Tests

The method presented in this study is completely new, and therefore its possible systematic errors have not been explored before. In particular the phase difference statistic has never been used in the contest of the Lyα forest, and given the unexpected results that we obtained it is necessary to carefully ponder all the possible effects that may change our conclusions and to revise our main assumptions. For sake of clarity, we classify the sources of uncertainty we could think of into four broad categories:

- *Data-originated* : calculating phases from the Lyα forest of observed pair is not a straightforward operation, partly for mathematical reasons (see § 4.2) and partly for the presence of contaminants and the uncertainty on the continuum emission;
- *Calibration errors* : we employ a forward-modeling approach in order to adapt our set of simulation to the observations. This process involves several steps and assumptions with which we could inadvertently introduce sources of bias;
- *Model assumptions* : we base our phase-distribution prediction on a grid of thermal models built on top of a dark matter simulation. This clearly involves strong assumptions on the distribution of gas in the IGM, raising doubts on the applicability of this method to real data;
- *Statistical approximation* : the present method makes use of statistical tools such as Gaussian process interpolation and MCMC, which are approximated algorithms whose accuracy need to be tested.

In the following we will explain the test we perform in order to check the robustness of our method against these potential source of biases and error. However we caution that the validity of several of these tests is limited to the current level of accuracy. As it is natural, when the amount and the quality of the data will permit measurements of percent-level precision, also the requirements on the theoretical understanding and on the modeling of biases will be tighter.

5.3.1 Data-Originated

5.3.1.1 Phase Calculation

In chapter 4 we presented two possible ways of calculating phases of irregularly sampled functions: one employs least-square spectral analysis (LSSA), the other consists in rebinning the function into a regular grid and subsequently applying the standard discrete

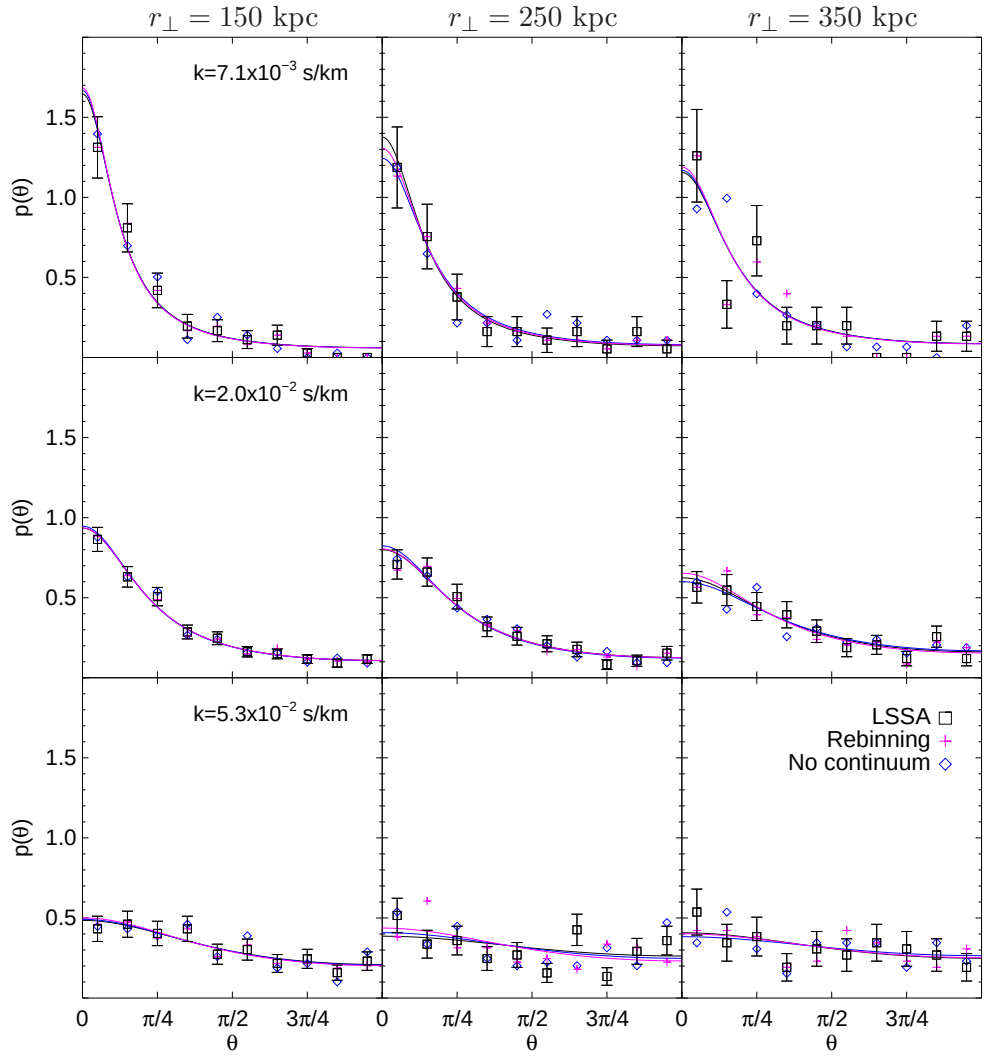

FIGURE 5.13: Phases of real data calculated at redshift $z = 2$ with three different methods: least square spectral analysis (black squares), rebinning on a regular grid and FFT (magenta crosses) and LSSA *without* continuum renormalization. (blue diamonds). The correspondent wrapped-Cauchy best fits are shown as solid lines, matched by color. We show the comparison for the same $r_\perp$ and k bins of figure 5.3. The three methods agree remarkably well in all cases, and the wrapped-Cauchy fits are essentially overlapping, implying that the phase distributions in the three cases are statistically equivalent. This proofs that the approximated method that we use to calculate phases are solid, and that the phase statistic is insensitive to uncertainties on continuum placement.

Fourier transformation. Since the two methods imply complementary approximation, checking that they lead to consistent phase distributions is a good check of the stability of this calculation. In figure 5.13 we show phases of the observed forest of quasar pairs binned as we have done for figure 5.3, adopting both the LSSA method (black squares) and the rebinning procedure (magenta crosses). In almost all cases the two methods agree extremely precisely, and most importantly the statistical estimator that we use in the likelihood, i.e. the wrapped-Cauchy concentration parameters, are practically identical. This can be seen by comparing the best-fit wrapped-Cauchy functions of the two distribution (black and magenta solid lines), which are indistinguishable at all $r_\perp$ and $k_{||}$. We also stress that the approximated Fourier transformation enters the forward-modeling of simulations, so even in the case where there was a significant effect on phase distributions, it would have been taken into account in our calibration.

5.3.1.2 Continuum Fitting

Among the properties of phases listed in chapter 3 we claimed that they are robust against uncertainties on continuum fitting. This was argued based on the mathematical definition of phases, which are invariant under a global renormalization of the function. If continuum error could be described as an uncertainty on the renormalization than our statement would be exact. In the realistic case of a fluctuating continuum, phase distributions are still untouched in an approximated sense if such continuum fluctuations occur on scales larger than the typical modes that we want to use ($v_{||} \gtrsim 1500$ km/s). This could not be true in the presence of associated lines, like BALs, or near the Lyα and Lyβ emission lines. As explained above, we exclude from the sample quasars with recognizable BALs, and we do not attempt to use the forest in the vicinity of the two emission lines. To proof explicitly that phases are not sensitive to continuum errors, we estimated phase distributions *without fitting the continuum* of the spectra, directly from the observed flux, and we compare them with the standard case of continuum-renormalized spectra. The results of the calculation are shown as blue diamonds in figure 5.13. The agreement is remarkable at all $r_\perp$ and $k_{||}$, both with the interpolation and LSSA methods. Even where there are differences on the actual distribution, the fitted wrapped-Cauchy function almost coincides with the continuum-corrected phases. This test proofs that phases are insensitive to variation of the continuum, unless for some reason other than BAL there are neglected fluctuations and wiggles at small velocity scales.

5.3.1.3 Contaminants

It is possible that part of the absorption in the Lyα forest of our spectra is not caused by neutral hydrogen in the diffuse IGM but from other systems that are not modeled in our simulation. An examples are broad absorption lines (BAL) associated with quasars, which can have high velocities and be blueshifted in the Lyα forest. As we have just reminded, all the QSOs that exhibit such lines have been removed from the sample. Similarly, we have excluded all the region of the forest where we could identify a Damped LyαAbsorber (DLA), since they are not described by the optically-thin approximation that we adopt. For the same reason also Lyman Limit Systems (LSS) should be removed, but this is not possible because they are practically indistinguishable from the forest. However, we doubt that they can generate a strong bias, given that they contribute to the forest absorption by only a tiny amount [McDonald et al., 2005]. A similar arguments holds for metal contamination. Moreover, metal lines are narrow, and they affect mostly Fourier modes with $k > 0.1$ s/km [McDonald et al., 2000] which we exclude precisely for this reason.

If metal contamination and LLSs have a stronger impact than expected, they might cause a decrease in the transverse coherence of pairs, since they would not be strongly correlated in space. This effect would lead to an underestimation of the Jeans filtering scale, henceforth we plan for the future a more careful and quantitative test of the robustness of our results with respect to these contaminants.

5.3.2 Calibration

5.3.2.1 Resolution

The first step of the forward modeling consists in convolving the simulated skewers with a Gaussian kernel whose width is regulated according to the estimated resolution. This operation does not change phase differences initially, but it cuts the longitudinal power exponentially at high k, which have a significant effect on phases when noise is added (see discussion in § 5.1.5). Moreover, the estimation of the resolution also set the maximum $k_{||}$ we will use in the phase likelihood for that pair. It is then natural to ask what kind of bias we would get if the resolution we are assuming is underestimated or overestimated. In order to explicitly check this, we perform a test using a mock data sample from our simulation at $z = 3$. We generate skewers with S/N=10 and resolution FWHM=100 km/s, which are average values for our quasar sample, chosen from a fiducial model with $\lambda_J = 80$ kpc. We then try to "measure" the Jeans scale of this mock sample applying our standard technique, but calibrating the simulations

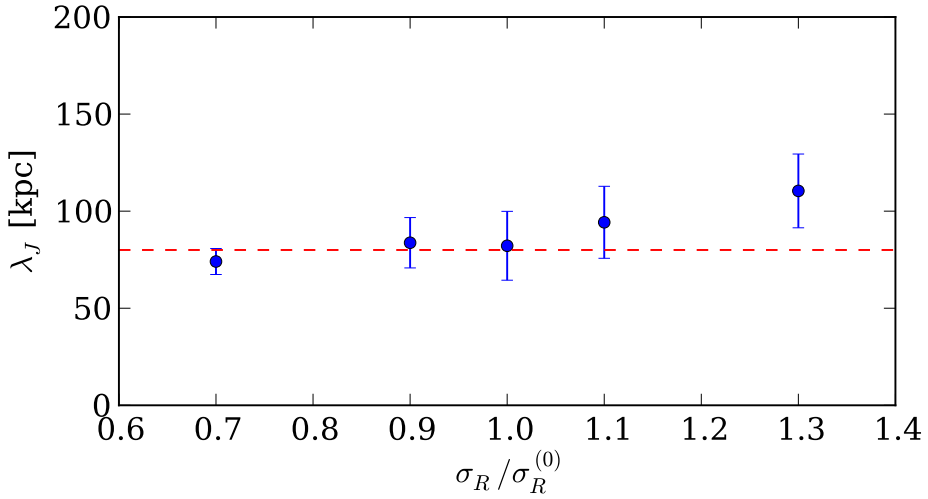

FIGURE 5.14: Bias on the Jeans scale measurement deriving from a wrong resolution estimation. We perform a "measurement" of the Jeans scale from a mock sample of skewer pairs taken from a fiducial model with $\lambda_J = 80$ kpc. The correct Jeans scale is marked by the dashed red line. The blue points represent the estimated Jeans scale as a function of the assumed resolution kernel width σ_R ≈FWHM/2.335 relative to the one of the mock data $\sigma_R^{(0)}$. This plot suggests that the result are stable especially for *under*estimation of σ_R (i.e. overestimation of the resolution), but a significant overestimation of λ_J is possible if the noise is overestimated by $\gtrsim 30\%$.

with the wrong resolution kernel. We test a 10% and 30% error on the resolution, both by underestimation and overestimation. For comparison, we also do the test with the correct FWHM. The results are shown in figure 5.14, where we plot the estimated Jeans scale against the assumed width of the resolution kernel σ_R, expressed relative to the correct one $\sigma_R^{(0)}$. The Jeans scale of the fiducial model is marked by the red dashed line. From this test we conclude that the measurement is relatively robust with respect to resolution uncertainties, although a significant bias would be caused by an overestimation of σ_R (i.e. an underestimation of the resolution) of the order of 30%.

As a further test of our resolution calibration, we split the data sample at $z = 2$ in two sets with high (FWHM$<$ 100 km/s) and low (FWHM$>$ 100 km/s) resolution and compare the separate constraints on the Jeans scale. If our calibration is correct for all the instruments used to observe the pairs, the results of a measurement from the two dataset should be consistent. Figure 5.15 shows that this is verified for our test, at least for the achievable degree of precision.

5.3.2.2 Skewer Extension

In section 4.3.3 we described how we extend the simulated skewers by replication in order to match the length of the observed Lyα forest chunks, arguing that this should

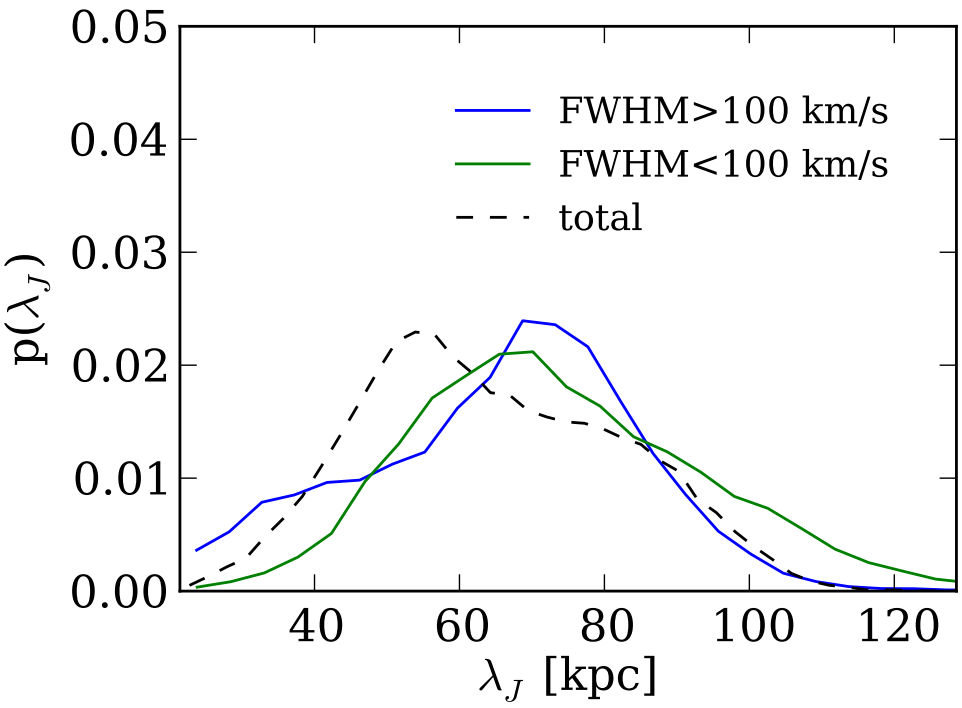

FIGURE 5.15: Posterior probability distribution for λ_J for the subsamples with resolution lower and higher than FWHM=100 km/s at $z = 2$. The constraints from the two subsets are fully consistent.

not create artifacts in phase distributions. We do a simple test to verify our statement, by applying the extension to simulated pairs and by comparing the final phase statistic with that of the unextended spectra. The results are shown in figure 5.16. We generate pairs with LRIS resolution of about FWHM= 150 km/s and a S/N of 10, building a mock sample at a chosen transverse separation and redshift. We do the test using a sample at $z = 2$ at $r = 132$ kpc (black lines), and one at redshift 3 and impact parameter $r = 432$ kpc (red lines). We then calculated the phase distributions at all k-bins and the relative wrapped-Cauchy ζ parameters. These are marked as a function of k by the thick lines in the figure, and represent the "original" phase distribution. We then extend the skewers as illustrated in 4.3.3 by a factor 3.2 and 2.6 in the $z = 2$ and $z = 3$ cases, respectively. Before calculating phases, we need to preliminary split the extended skewers into chunks of the length of the original box, in order to preserve the Fourier bases. We then calculate phases separately in each chunk and use the ensemble obtained in this way to calculate the ζ parameters. These are shown as thin lines, and agree in both cases with those of the original box, implying that the phase statistic is correctly preserved. The dotted-dashed lines refer to the case where the preliminary chunking has been neglected, and clearly show how this would cause an artificial decrease in coherence.

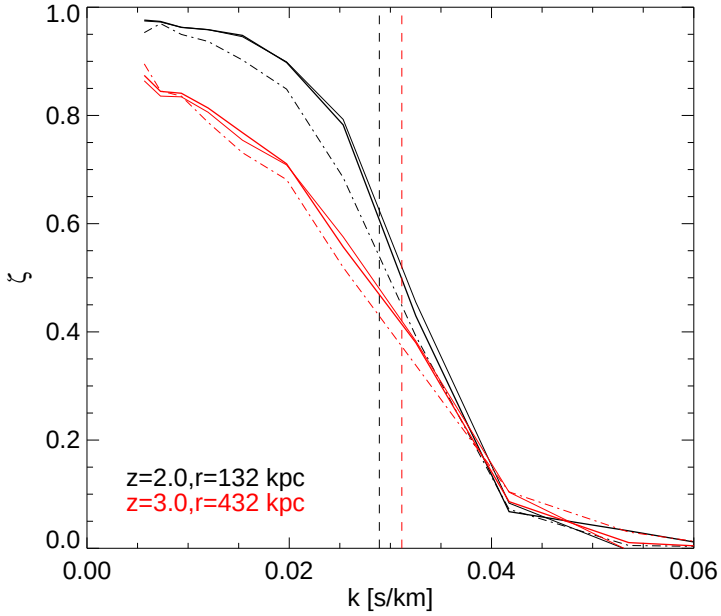

FIGURE 5.16: Effect of the periodical extension of skewers in the simulation. We calculate the Cauchy ζ parameters of phase differences for all the k-bins in two cases. The black lines represent a set of skewers from the snapshot at $z = 2$ and at $r_\perp = 132$ kpc, where the grid has been extended 3.25 times. The red lines are calculated at $z = 3$, from skewers at separation of 432 kpc and extended by a factor 6.1. A S/N of 10 per pixel has been assumed in both case, with a resolution limit (correspondent to LRIS) marked by the vertical dashed lines (the discrepancy between the resolution limits is due to the different wavelength range at different z). The Cauchy parameters calculated after extension (thin lines) are fully consistent with those of the unextended box (thick lines), meaning that the extension procedure does not alter significantly the statistical distribution of phases, as long as we follow the correct procedure of separately calculate phases on chunk of the same size of the original box. The dotted-dashed line shows the error one could commit by extracting phases directly from the extended grid, without the preliminary chunking.

5.3.2.3 Noise

Noise is taken into account in our model by adding random fluctuations to the simulated spectra according to the estimated error σ_N. Since the coherence of phases is significantly decreased by noise, it is important to test what bias might arise in the case of a inaccurate estimation of such errors. We do an analogous test to what we have done in § 5.3.2.1 to test the robustness to resolution estimation. We used the same mock sample of pairs with S/N=10 and resolution FWHM=100 km/s from the fiducial model with $\lambda_J =$ 80 kpc. This time we repeat the "measurement" on the mocks varying the assumed noise level. We calibrate skewers by adding Gaussian noise with standard deviation σ_N underestimated or overestimated by 10% and 30% compared to the exact width $\sigma_N^{(0)}$, and we limit the dynamic range to $k < 1/\sigma_R$. The results are shown in figure 5.17, where the correct Jeans scale is marked as a red dashed line. The bias deriving from a wrong assumption on the noise is stronger than a comparable error on the resolution. In

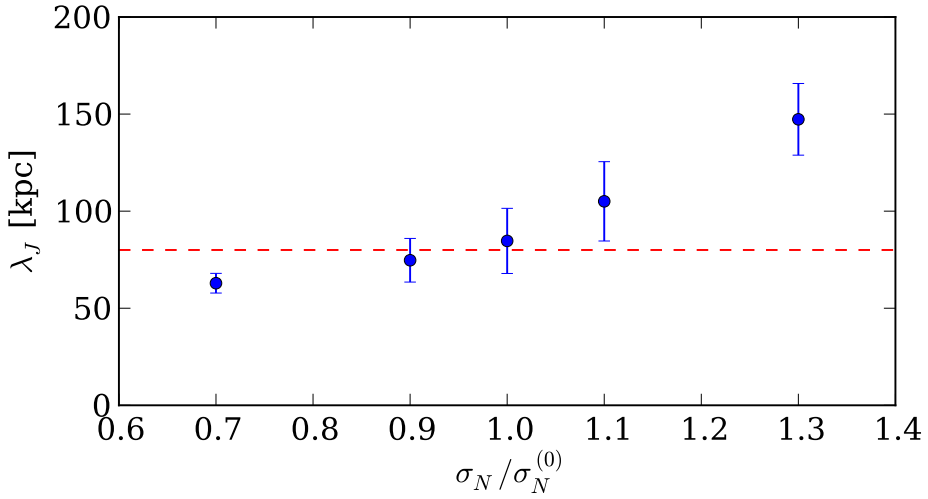

FIGURE 5.17: Bias on the Jeans scale measurement deriving from a wrong noise estimation. We perform a "measurement" of the Jeans scale from a mock sample of skewer pairs taken from a fiducial model with $\lambda_J = 80$ kpc. The correct Jeans scale is marked by the dashed red line. The blue points represent the estimated Jeans scale as a function of the assumed noise level σ_N relative to the exact noise level of the mock data $\sigma_N^{(0)}$. This plot suggests that the result are relatively stable for *under*estimation of noise, but a significant overestimation of λ_J is possible if the noise is overestimated by $\gtrsim 10\%$.

particular an overestimation of the noise would lead to a significant *overestimation* of the Jeans scale, up to almost a factor 100% in the worst case of a 30% overestimation of σ_N. Interestingly, the *underestimation* of the Jeans scale due to an underprediction of noise is much weaker. Similarly to what we have done in § 5.3.2.1, we test the consistency of our noise modeling by splitting the $z = 2$ sample in high-quality (S/N> 20 per Angström) and low-quality (S/N< 20 per Angström) pairs, where the S/N of a pair is defined according to eq. 5.6. The comparison of the constraints from the two subsamples shows some tension, but not at a statistically significant level.

5.3.3 Model Assumptions

We obtain the prediction of phase differences in function of λ_J from a simplified IGM model built on a DM simulation. In the next chapter we will test this model using hydrodynamical simulations to verify its accuracy. Here we check its internal consistency by exploring the sensitivity of different parts of the dynamic range. If the models we use are a sensible representation of the IGM, pair at different impact parameter $r_\perp$ and modes at different wavenumbers $k_{||}$ should provide compatible constraints on the physical parameters. In the first test, we perform the measurement of the Jeans scale separately on close ($r_\perp < 200$ kpc) and wide ($r_\perp > 200$ kpc) pairs at $z = 2$ and we subsequently compare the inferred posterior distributions for λ_J. The results are shown

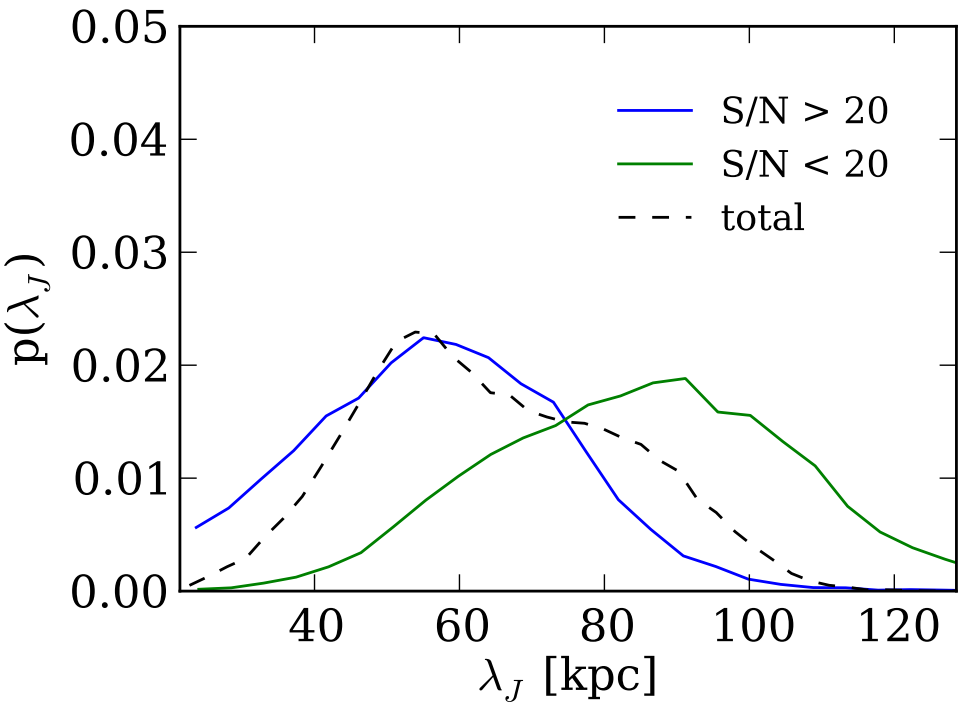

FIGURE 5.18: Posterior probability distribution for λ_J for the subsamples with signal-to-noise ratio lower and higher than 20 (per Angstrom). There is a slight tendency of higher-S/N data of pointing towards lower Jeans scales, but the two distributions are statistically consistent, suggesting that there are no significant biases due to S/N estimation at the current level of accuracy.

in figure 5.19. The plot visually suggests close pairs tend to favor smaller value of λ_J, but the two distributions are still statistically consistent. As a second test, we do a similar analysis by using separately phases relative to low-k modes ($k_{||} < 0.017$ s/km) and phases at high-k ($k_{||} > 0.017$). The result is shown in figure 5.20, and indicates good agreement between the two subsamples.

5.3.4 Statistical Approximations

5.3.4.1 Wrapped-Cauchy Distribution

The likelihood function employed in our Bayesian analysis assumes that the predicted phase distributions are precisely described by the wrapped-Cauchy function. This is particular convenient since it allows us to compress the information of the full phase PDF into a single number (the ζ-parameter) at each $r_\perp$ and $k_{||}$, but it could be a source of bias if the phase probabilities are not faithfully traced by the wrapped-Cauchy fit. In figure 5.21 we demonstrate the level of agreement that we get between the full simulated distributions and our fits. In this plot we have assumed a combination of

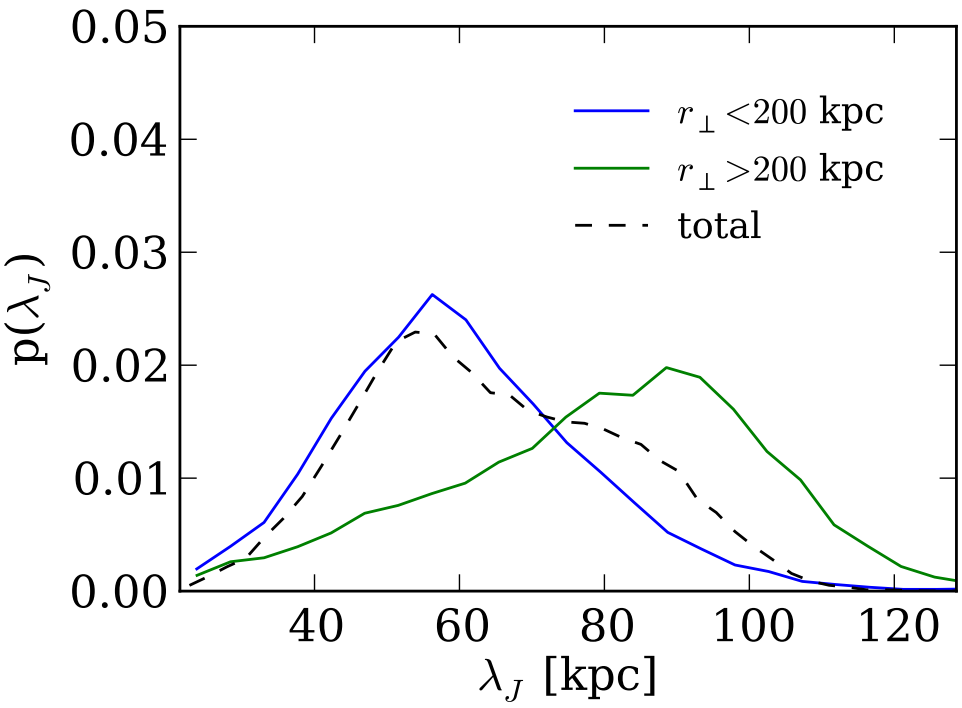

FIGURE 5.19: Posterior probability distributions for λ_J for the subsamples of pairs at separation larger and smaller than 200 kpc. Data from closer pairs seems to favor slightly lower Jeans scales, but also in this case the two distributions are statistically consistent.

resolution and S/N correspondent to our data sample, analogous to what we have done in figure 5.3. The differences between the wrapped-Cauchy fits (solid lines) and the actual distributions (diamonds) are almost unnoticeable for all the values of $\lambda_j, r_\perp$ and k that we show in the figure.

5.3.4.2 Emulator

We explained in § 5.1.4 that our emulator is reliable if the training grid is dense enough with respect to the smoothness of the interpolated function in the parameter space. When this condition is not fulfilled our results could be sensitive to the parameters we choose in implementing the Gaussian-process interpolation. These parameters are the *smoothing lengths*, which express the degree of correlation we assume when interpolating in parameter space. Choosing large smoothing lengths prevent the interpolated variable to vary strongly between the grid points, while small smoothing lengths cause the prediction to fall quickly to zero when no neighbor points are present. To check that our results are independent of this choice we repeat the measurement after varying the smoothing length by ±20%. We then looked at the confidence levels and verify that they

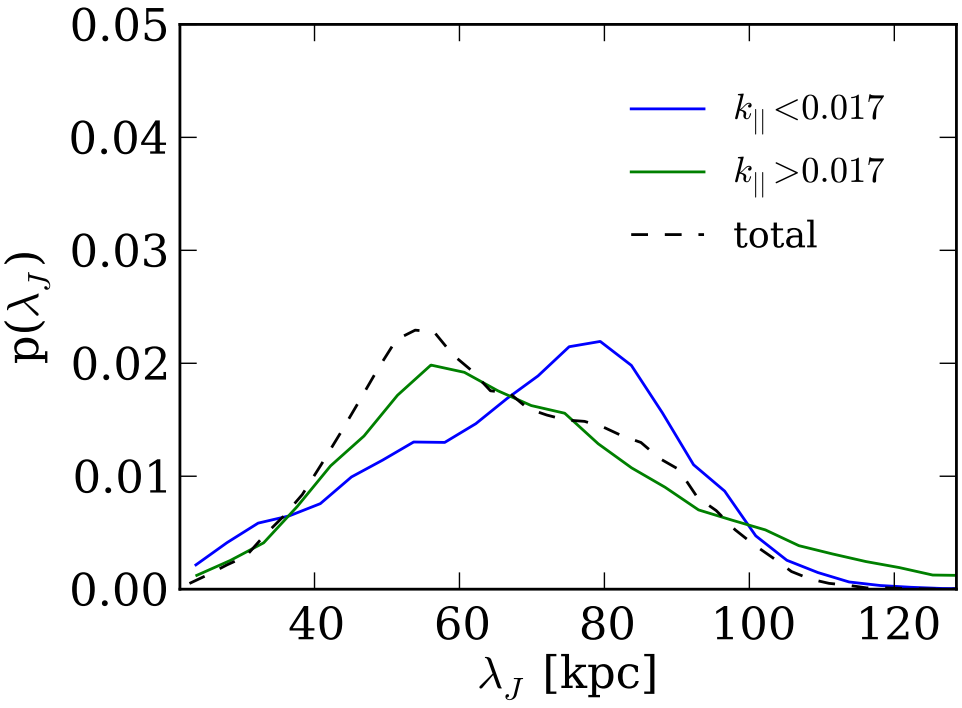

FIGURE 5.20: Posterior probability distributions for λ_J splitting the phases by wavenumber. The constraints originated from the phases with $k_{||} < 0.017$ s/km are statistically consistent with those obtained from phases at $k_{||} > 0.017$ s/km.

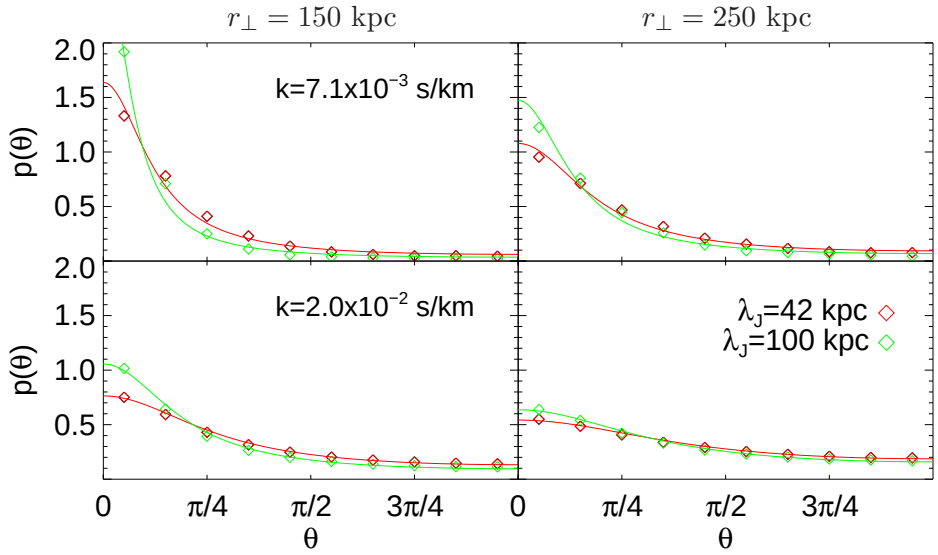

FIGURE 5.21: Example of wrapped-Cauchy fits to the distributions predicted by our simulations. This figure show the phase PDFs at $z = 2$ for $\lambda_J = 42$ kpc (red) and $\lambda_J = 100$ kpc (green). The panels refer to the impact-parameter bins $[100, 200]$ kpc and $[200, 300]$ kpc (left to right) and the k-bins $[0.005, 0.01]$ s/km and $[0.01, 0.04]$ s/km (top to bottom). The wrapped-Cauchy function traces the actual distributions with excellent agreement in all cases.

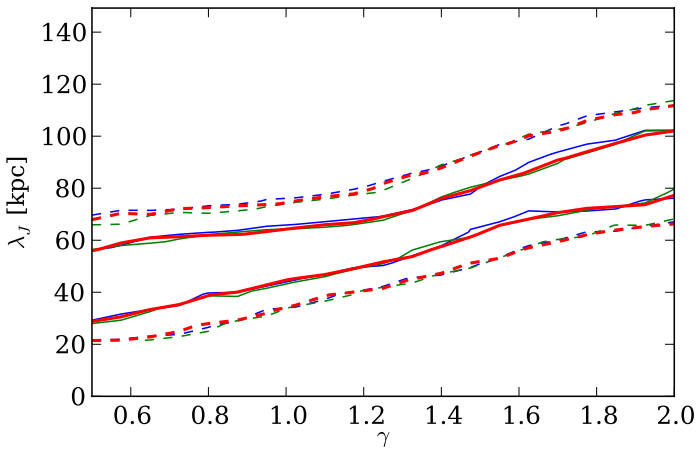

FIGURE 5.22: Convergence of the emulator. We plot the 65% (solid lines) and 95% (dashed lines) confidence level in the λ_J-γ plane, for three choices of the *smoothing lengths*. The default choice is represented by the red lines, while the blue/green contours are obtained by assuming 20% larger/smaller smoothing lengths. The lines trace each other with high accuracy, implying that the emulator interpolation does not affect our results.

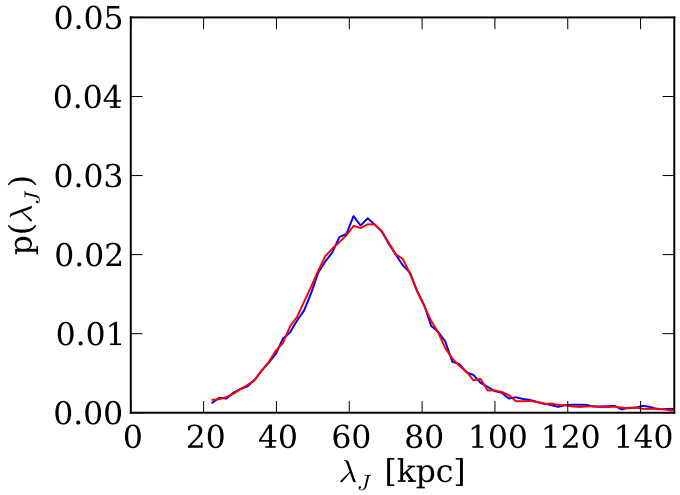

FIGURE 5.23: Convergence of the MCMC chains for the posterior at $z = 3$. The histograms are calculated from chains with 180000 (blue) and 45000 (red) points.

are converged. We show this test for the most delicate case of the λ_J-γ plane at $z = 2$, where the contours are narrower and thus have the most demanding requirement on the density of the training grid. Figure 5.22 demonstrates that the emulator is converged at much higher precision than the accuracy of the measurement.

5.3.4.3 MCMC convergence

We check that our MCMCs are converged by running a longer chains and comparing the posterior distribution for λ_J. The calculation of the likelihood is relatively fast, since it only requires the interpolation of one number (the logarithm of the likelihood on the grid), and our parameter space has only 3 dimensions. Moreover, the likelihood function is relatively smooth in the parameter space at the current level of precision. All these factors favor a fast convergence of MCMC runs, as it is shown in figure 5.23.

Chapter 6

Interpretation and Discussion

In the previous chapter we presented our estimation of the IGM Jeans scale achieved by calibrating phase differences on a set of models based on a dark matter N-body simulation. In the context of our model, λ_J is the width of the kernel with which we smooth the dark matter density field in order to obtain the baryon distribution. In a CDM universe, where dark matter has no smoothing length, this identifies λ_J with the smoothing length of baryons, or thinking in Fourier space, as the scale where the 3d matter power spectrum is truncated. This interpretation however only holds if the assumptions made to build such models are a good approximation of the real IGM. We need to understand if this smoothing length exists in the universe and if we are able to probe it using our phase-difference method.

In this chapter we try to address this problem by means of full hydrodynamical simulations (which I briefly present in § sims). By doing so we complete the task left open in chapter 4 of demonstrating the validity of our method, and we also gain useful understanding on the meaning of the measured λ_J. The key idea is that the Jeans filtering can be identified once the appropriate density range is selected: in general density peaks of collapsed objects dominate the matter power spectrum at small scales, and the suppression due to pressure is completely concealed. Once the high densities are removed from the analysis the power spectrum of the IGM emerges and the truncation due to pressure support clearly appears. The fact the Lyα forest is only sensitive to the low-density regions motivates this reasoning from an observational point of view, indicating this filtering scale as a natural interpretation of the quantity we measure with quasar pairs.

The properties of the Jeans filtering scales are still poorly understood, in particular its sensitivity to the thermal history and the expected value at the Lyα-forest redshifts. These and other related theoretical questions are now under research in our group, an

effort in which I have been directly involved in the last part of my PhD. I will illustrate in the second part of this chapter some preliminary results that we obtained in this direction, in particular the possible ways of defining the filtering scale in hydrodynamical simulations (§ 6.3), and finally the comparison of the measured filtering scales with the expected values for a set of thermal histories in § 6.4.

6.1 Hydrodynamical Simulations

I this section I will refer to a set of hydrodynamic simulations that have been run for multiple IGM-related science goals. We used both the Lagrangian code Gadget3, an improved version of the publicly available code Gadget2 [Springel, 2005], and the recently developed Eulerian code Nyx [Almgren et al., 2013].

Gadget3 was run with 2×512^3 gas and dark matter particles in a 10 Mpc/h box. To optimize the calculation, we used the "quick Lyα" flag, that converts gas into stars above an overdensity threshold (in our case $\Delta > 1000$). This method does not affect the IGM and speeds up significantly the simulation.

Nyx simulations are run on 10 Mpc/h size cube with 512^3 cells. This boxsize and resolutions are chosen in order to achieve convergence in the low-density IGM and in particular of the phase-difference statistic of quasar pairs (Oñorbe et al.,in prep.)

In both simulations we assume the gas to be optically thin and in ionization equilibrium with a spatially homogeneous ultraviolet background (UVB). We adopt the UVB from the the model of Haardt & Madau [2012]. In order to study different thermal history we follow the procedure used in Becker et al. [2011]. The photoheating rates for HI,HeI and HeII are rescaled in a density-dependent manner as $\epsilon = A\Delta^B\epsilon_0$, where $\Delta = \rho/\bar{\rho}$ is the overdensity and ϵ_0 the Haardt and Madau photoheating rates. A detailed description of the simulation will be given in Oñorbe et al. (in prep.) and Kulkarni et al. (in prep.). For what concerns this chapter, we will use the results from the Nyx simulation with $B = 0$ and $A = 1, 0.5, 0.1$, to which we will refer as NHM, N0.5HM and N0.1HM respectively, and from Gadget 3 using $B = 0$ and $A = 1$ (GHM).

In these simulation we use the cosmological parameter $\Omega_m = 0.275, \Omega_b = 0.046, \Omega_\Lambda = 0.725, h = 0.702$ and $\sigma_8 = 0.816$. The other details of the four simulations are specified in table 6.1.

Name	Code	N_p	L [Mpc/h]	A	B	T_0 [K]	γ	$\lambda_{J,\mathrm{fit}}$ [kpc]
NHM	Nyx	2×512^3	10	1	0	10919	1.56	82
N0.5HM	Nyx	2×512^3	10	0.5	0	7029	1.56	67
N0.1HM	Nyx	2×512^3	10	0.1	0	2504	1.58	46
GHM	Gadget3	2×512^3	10	1	0	9507	1.59	77

TABLE 6.1: List of the simulations discussed in this chapter. A and B are the parameter regulating photoheating rate of the IGM $\epsilon = A\Delta^B\epsilon_0$, where ϵ_0 are the photoheating rates of the Hardt and Madau model [Haardt & Madau, 2012]. $\lambda_{J,\mathrm{fit}}$ is the value of the Jeans scale obtained from the fit of the real-flux power specutrm (see 6.3). The values of T_0, γ and λ_J refer to the snapshot at $z = 3$. The parameters of the temperature-density relation T_0 and γ are obtained by fitting the (volume-weighted) probability distribution in the T-Δ space from the simulations.

6.2 The Filtering Scale in the *Real-Flux* Field

Our method to measure the Jeans scale relies on a set of simplified models of the IGM, based on the particle distribution of Dark-Matter simulations. In particular, we are assuming that baryons faithfully trace dark matter density, with the only difference of a characteristic smoothing length set by the pressure. The second strong hypothesis is that the smoothing scale λ_J is a constant value across the volume, independent on the temperature and on the density. We acknowledge that in this way we neglect a number of relevant physical processes which would require hydrodynamics and radiative transfer to be correctly taken into account. Moreover, treating the Jeans scale as a fixed quantity is unphysical, given that it is expected to scale as $\lambda_J \propto \sqrt{T/\rho}$ (although the effect of thermal history at different densities is not clear).

The technical difficulties in running large grid of models with full treatment of hydrodynamics are the main justification of our approach, but one may wonder not only whether our measurement is reliable, but also if the whole problem has a well-defined physical meaning. If the Jeans scale is dependent on the local physical parameters, there would not be any *global* Jeans scale in nature, and our attempt to measure it might thus appear pointless.

In this section we tackle these two questions, clarifying what is the physical meaning of our measurement and that under this perspective our simplified DM models are sufficiently accurate.

6.2.1 Is there any Jeans Scale of the IGM?

The classical argument defines the Jeans scale in the context of linear theory, where density and temperature are effectively constant over the space. Under this conditions, the Jeans scale is also constant and can be regarded as a global parameter. However, the same definition does not apply to the real universe at the typical redshifts of the Lyα forest ($2 < z < 4$), for the simple reason that both temperature and densities are not expected to be homogeneous. In particular the values of the Jeans scale estimated in our measurement are well below the nonlinear scale at this epoch. This degree of complexity requires a generalization of the definition of λ_J. One possibility is simply to consider it as a local quantity and preserve the classical definition based on the local physical condition of density and temperature. As discussed in chapter 1, this choice does not have a clear physical meaning in the context of the IGM, where the dynamical time and the sound-crossing time λ_J/c_s are of the order of the Hubble time, and the combined effect of expansion and thermal history must be taken into account. Moreover, it would be our task more difficult since we would need to predict a full distribution of Jeans scales, instead of a single values.

The most natural approach for our purposes is to extend the definition of λ_J as a geometric property of the density, i.e. as the smoothing length of the baryonic component, commonly known as *filtering scale*. Such smoothing length can be defined starting from the correlation function or from the power spectrum, where we expect a cut-off at the correspondent scale. From now on we will only reason in terms of the power spectrum in Fourier space, consistently with the rest of this work.

Providing a precise definition it is not a straightforward operation for several reasons. First of all, if the filtering scale depends on pressure it has to share with the classical Jeans scale the property of being density-dependent, in a way which is however not understood, neither from theory nor from observations. We need to define a sort of *effective* filtering scales across densities and temperature, but it is not obvious that such a scale should emerge when the full density range is considered. Secondly, if such scale exists, one needs to make assumptions on the intrinsic shape of the power spectrum and on the filtering function in order to associate a scale to an observed cutoff, as will be discussed in § 6.3.

Figure 6.1 answers the question raised above: the 3d power spectrum of baryons ($\Delta < 1000$) does note exhibit any cutoff. This can be view as a consequence of the contribution of collapsed objects to the power at small scales. This contribution, although dominant at high k, is not relevant in the context of IGM studies as collapsed structures occupy only a tiny portion of the volume, despite containing a significant fraction of the mass.

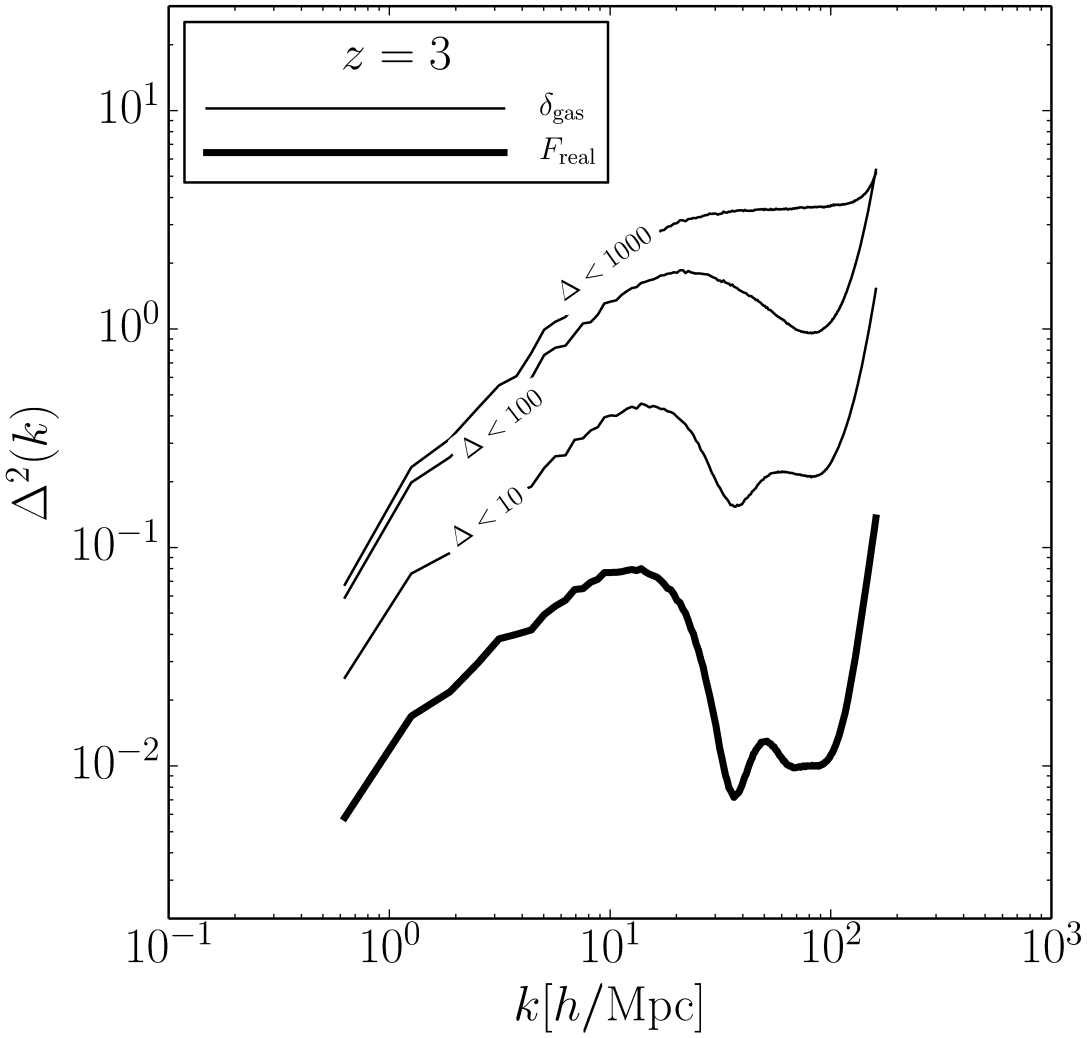

FIGURE 6.1: From Kulkarni et al.(in prep.). The plot shows the 3d dimensionless power spectrum of gas (thin lines) and the real flux (thick line) in the GHM simulation. The three gas power spectra are calculated assuming three different density threshold $\Delta_{th} = 1000, 100$ and 10. The highest threshold correspond to the density at which the code transforms gas into stars. As the threshold decreases lower densities are selected, and the Jeans cutoff emerges. Analogously, the transmitted Lyα flux in real space naturally suppresses the contribution of high densities, and its power spectrum is therefore strongly dependent on the effect of pressure. The upturn at $k > 10^2 h/{\rm Mpc}$ is due to a simulation artifact.

Most important they are not probed by the Lyα forest since they gives raise to completely saturated lines. It is convenient to remove them from the analysis by clipping the matter field, i.e. setting an upper threshold to the overdensity Δ_{th}. This consists in the simple transformation

$$\Delta_c = \begin{cases} \Delta & \text{if } \Delta \leq \Delta_{th} \\ 0 & \text{if } \Delta > \Delta_{th} \end{cases} . \tag{6.1}$$

Figure 6.1 shows how the cutoff appears and evolves as the overdensity limit Δ_{th} decreases, corresponding to an increased suppression of structures. This behavior is expected for two reasons: the contribution of collapsed regions vanishes and pressure support is more effective at low density, where gravity is weaker. It is natural to consider the location of this cutoff as the filtering scale of the IGM, however the fact that it shifts with Δ_{th} brings us back to our initial problem of defining unambiguously the filtering scale.

The Lyα forest offers a very natural solution to this problem. In a broad sense, the transmitted flux can be considered as a clipped version of the density field, where the suppression of high densities is originated in the exponentiation $F = \exp(-\tau)$ and not as a sharp threshold, but it is equally effective in removing the contribution of high densities and isolating the property of the IGM. This suggests the possibility of applying the same type of transformation to the density field of our simulation and define the IGM filtering scale based on the geometrical properties of this field, in particular on the location of its 3d power spectrum. An important caveat is represented by the redshift-space nature of the Lyα forest: the absorption is distorted along the sightline by thermal broadening and the motion of the gas, so it does not faithfully trace the underlying matter distribution and it is not an isotropic 3d field. We circumvent this problem by defining the *real-space* transmitted flux F_r, i.e. the optical depth to Lyα absorption at each point in real space. This quantity is equivalent to the Lyα forest absorption in the limit of a cold and steady IGM, and it is given by the Fluctuating Gunn-Petersonn approximation [Gunn & Peterson, 1965]:

$$F_r = \exp\left[-\frac{\pi e^2}{m_e c} f_\alpha \lambda_\alpha H^{-1}(z) n_{HI}\right] \tag{6.2}$$

where f_α is the oscillator strength of the Lyα line and n_{HI} the density of neutral hydrogen. Studying this quantity has several advantages:

- it is by construction sensitive to the properties of the IGM and independent on the physics of galaxies and high-density regions in general;
- the way it weights different densities is neither ambiguous nor arbitrary;
- differently than the velocity-space flux, it has no sensitivity to thermal broadening, which would introduce degeneracies in the power spectrum and in particular on the cutoff, as we have discussed in detail in chapter 2
- it is closely connected to the observed Lyα forest. Although relating the latter with the real-space flux requires a precise modeling of peculiar motions and temperature, we have shown that the phase difference statistic is practically independent on them,suggesting that F_r can be directly probed using quasar pairs.

This properties lead us to conclude that the filtering scale obtained through our measurement should be identified with the filtering scale of the real-space flux 3d field, as it will be explained in the next section.

A minor disadvantage is due to the evolution of the Lyα absorption with redshift: as the forest becomes more transparent due to expansion the range of densities probed by F_r moves to higher values, making less intuitive the comparison between filtering scales at different epochs.

6.2.2 Validation of the Dark-Matter Models

The identification of the Jeans scale λ_J of our DM models with the cutoff of the power spectrum of F_r is motivated following a logic analogous to the previous section. By construction, the filtering scale of the density field is a constant parameter in our models. As a consequence, the 3d power spectrum of the density exhibit a cutoff whose location is invariant under the choice of different density ranges. As shown in the previous section, this is very different than what we expect to find in the IGM and what we see in hydrodynamical simulations. We must then explicitly state what is the equivalent of λ_J in the real universe, i.e. what is the physical meaning of our measurement. Equivalently, we must specify at which density we expect our estimate for λ_J to match the filtering scale of the IGM. The fact that the statistic we are using, phase differences, is measured on the Lyα forest and is insensitive to thermal broadening naturally leads us to conclude that F_r defines the desired range of densities.

To be precise, F_r is the transformed of the density field that we expect to have the smoothing length to which our method is sensitive to. Since F_r does not select exactly *one* density, we could still expect a variable filtering scale. If this variability is significant, our fixed-λ_J approximation may be too imprecise. This is one of the reasons why we need to test the method with simulations, but we can use the classic Jeans formula to get a broad intuition of the dispersion of the Jeans scale in the Lyα forest. Using a temperature-density relationship $T \propto \rho^{\gamma-1}$ and assuming $\lambda_J \propto \sqrt{T/\rho}$, one gets $\lambda_J \propto \rho^{\beta}$, with $\beta = \gamma/2 - 1$. For typical values of $\gamma \sim 1.2 - 1.6$, this dependence is quite shallow. Figure 6.1 shows that in any case an *effective* cutoff is clearly present in the power spectrum of F_r in the simulation GHM. If we could proof that,despite of the approximated models, our method is able to correctly predict the position of such cutoff, than we would demonstrate its validity, also showing that the interpretation we propose is correct.

We do this by "measuring" the filtering scale in the hydrodynamical simulation GHM at redshift 3 and by comparing the outcome with the cutoff of the 3d real-flux power spectrum calculated from the simulation. Our test consists in the following:

- We draw synthetic pairs of skewers from the hydrodynamical simulation at various separations, defining our mock (noiseless) dataset.
- We calculate phase differences for all the pairs in the mock sample, in the same dynamical range utilized in data ($k < 0.1$ s/km).
- We evaluate the likelihood of the obtained set of phases using the probability distributions predicted by our grid of DM-based models.
- The filtering scale of the simulation is defined as the Jeans scale λ_J of the maximum-likelihood DM model.
- Finally, we compare the 3d power spectrum of the real flux F_r for the hydrodynamical simulation with that of the best-likelihood DM model. We remind that this implies smoothing the DM particle distribution using a pseudo-Gaussian kernel with $\sigma = \lambda_J$ and applying the FGPA to the field obtained in this way.

We performed this test on the GHM simulation (Figure 6.2) and on the NHM run (Figure 6.3). Although the general shape of the power spectra generally differ between hydros and our models, the location of the cutoff is remarkably well aligned in both cases. The Jeans scales estimated for the two simulations are approximately 110 and 128 kpc for GHM and the NHM run, respectively. This test confirms our hypothesis on the nature of the filtering scale measured via phase differences, and shows that no bias arises from the approximations assumed on the DM models which we use to calibrate phase distributions, at least at the current level of accuracy ($\sim 20\%$).

6.3 Definition of the Jeans Scale in Hydrodynamical Simulations

In the previous section we have argued that the Jeans filtering scale should be defined based on the cutoff of the 3d power spectrum of the real-flux field F_r. We have also shown that the location of this cutoff is what the phase difference statistic is mostly sensitive to. However, we have not provided yet a quantitative expression that defines λ_J in simulations and allows a direct comparison with the measurement. The most

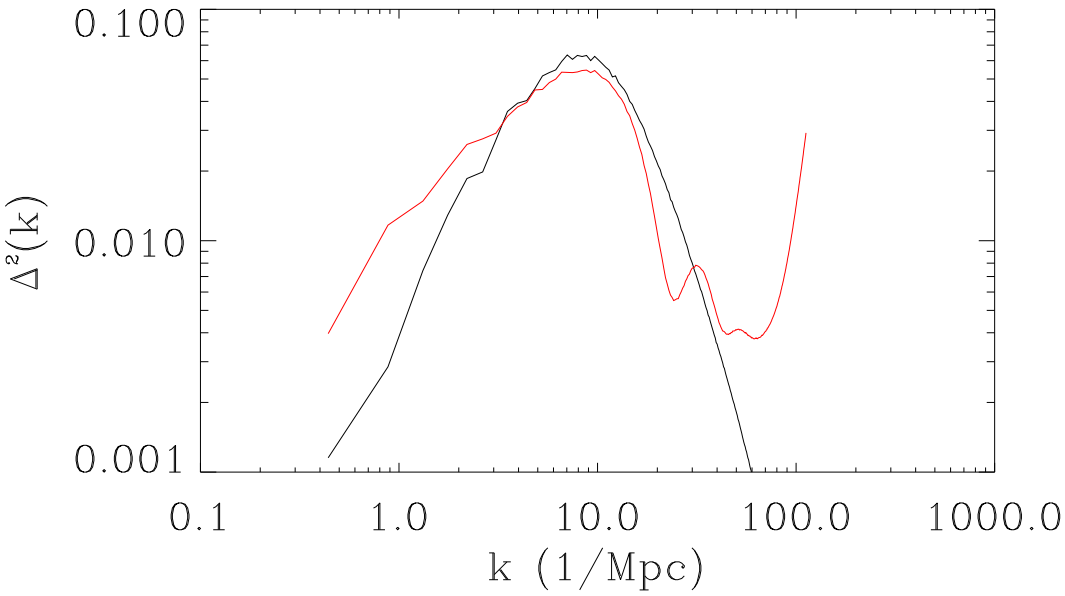

FIGURE 6.2: Comparison of the 3d flux power spectrum as calculated from the simulation GHM (red) and the one predicted by a "measurement" conducted with our method on a set of pairs extracted from the hydro run (black). The dashed vertical line explicitly mark the expected position of the cutoff, $k_J = 1/\lambda_J$.

simple way of defining the cutoff is parametrizing the real-flux power spectrum as a truncated power law

$$P(k) = Ak^n \exp[-(k\lambda_{J,\mathrm{fit}})^2] \tag{6.3}$$

where the normalization A, the index n and the filtering scale $\lambda_{J,\mathrm{fit}}$ are the fitted parameter. Note that with the Gaussian-truncation assumption, the relation between filtering scale and cutoff is $k_J = 1/\lambda_J$ and not $k_J = 2\pi/\lambda_J$. Figure 6.3 shows how well this fit (dashed red line) follows the true power spectrum of F_r in the $z = 3$ snapshot of the NHM run, corroborating the ansatz of a truncated power law. The fit gives a value of $\lambda_{J,\mathrm{fit}} = 82.4$ kpc, significantly lower than what we obtain by "measuring" the jeans scale with phase difference of pairs of skewers from the same simulation, which is $\lambda_{J,\mathrm{pairs}} = 128$ kpc.

It is likely that this discrepancy is due to the different slope of the power spectra of the two models (hydrodynamic and DM-based) at low-k, which alters the definition of the cutoff. We stress that boxes have slightly different cosmologies, and they are only 10 Mpc/h large, therefore low-k modes could be affected by cosmic variance. A rigorous and consistent definition of the filtering scale in hydrodynamical simulations is still under research, as well as its relation to that measured with our current quasar-pair method. For the moment we adopt a phenomenological approach and we apply a correction to translate the Jeans scale $\lambda_{J,\mathrm{fit}}$ obtained from the fit to the 3d real-flux power spectrum

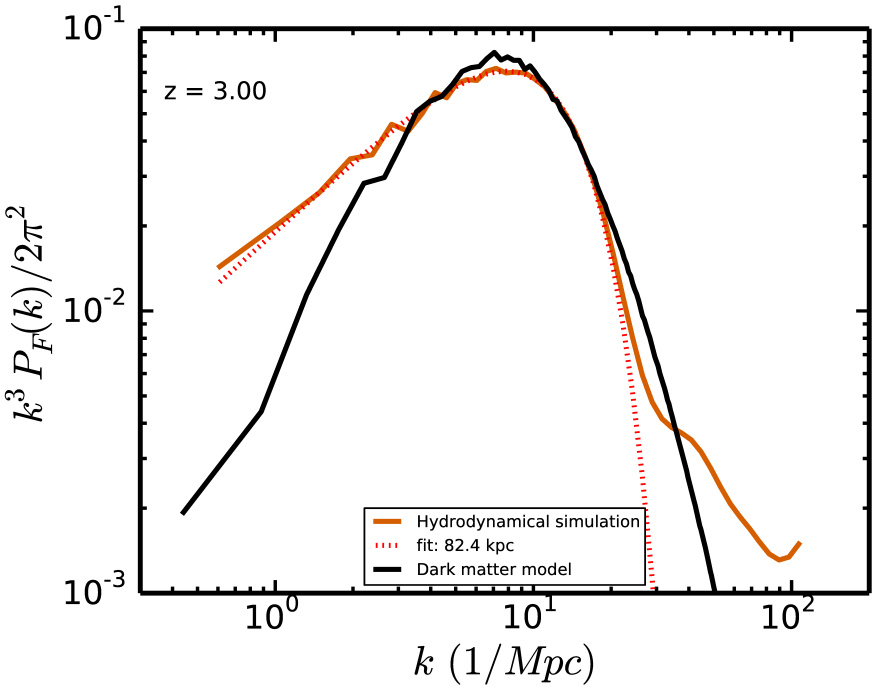

FIGURE 6.3: From Oñorbe et al., in prep. The plot shows the 3d dimensionless power spectrum of the real flux for the hydrodynamical simulation NHM (red solid line) and for the model based on the dark matter distribution smoothed with the filtering scale $\lambda_{J,\mathrm{pairs}}$. Here, $\lambda_{J,\mathrm{pairs}}$ is the Jeans scale measured from a set of synthetic pairs extracted from the same hydro simulation, using the same method we applied to data. The red dotted line is the fit to the red solid line obtained assuming a truncated power law $P(k) = Ak^n \exp[-(k\lambda_{J,\mathrm{fit}})^2]$. The position of the cutoff is again well matched, although the values of the parameters $\lambda_{J,\mathrm{pairs}} = 128$ kpc and $\lambda_{J,\mathrm{fit}} = 82.4$ kpc differ.

in simulations with the one estimated with the pairs method $\lambda_{J,\mathrm{pairs}}$. We assume this correction to be a multiplicative factor, which we tune on the snapshot at $z = 3$ of NHM (figure 6.3), giving

$$\lambda_{J,\mathrm{pairs}} \approx 1.4\lambda_{J,\mathrm{fit}}. \tag{6.4}$$

We will apply this relation in the next section in order to compare the results of our measurement with the prediction of the set of hydrodynamic simulation.

6.4 Redshift Evolution and Comparison with Simulation

Little attention has been devoted in the past in defining and predicting the filtering scale of the IGM. For this reason we are not yet in the condition of stating whether or not the results of our measurement meet the theoretical expectations of the typical IGM models. With the recent work described in the previous sections, however, we developed a consistent and physically motivated definition which allows us to draw the first conclusions.

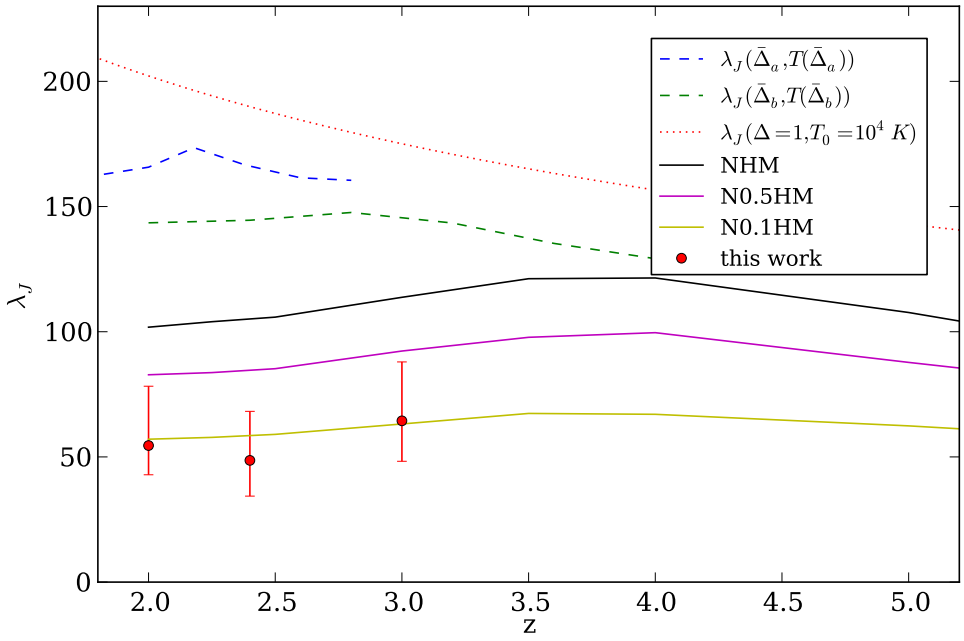

FIGURE 6.4: Evolution of the Jeans scale as measured from the sample of observed quasar pair (red dots) and predicted from different models. The red dotted line use the classical definition of the Jeans scale as a function of temperature and density, assuming a temperature of 10^4 K and a density equal to the mean of the universe. The green and the blue dashed lines also use the classic Jeans formula, but they are calculated using the typical density of the Lyα forest $\bar{\Delta}$ and the relative temperature $T_{\bar{\Delta}}$, as estimated in Becker et al. [2011] and Boera et al. [2014], respectively. The solid lines represent the filtering scale calculated in the simulations NHM,N0.5HM and N0.1HM (from top to bottom) by fitting the 3d power spectrum of the real flux and applying the correcting factor of 1.4 (see § 6.3 for details).

As a start, it is useful to consider the values expected if we identified the filtering scale with the instantaneous Jeans scale $\lambda_J^{(0)} = \sqrt{c_s^2/4\pi G\rho}(1+z)$, where the factor 4π in the denominator derives from the Gaussian truncation hypothesis. In figure 6.4 we plot in red (solid) the value of the classic Jeans scale at the mean density, assuming a constant temperature of $T = 10^4$ K. This curve is clearly too high to be consistent with our measurement (red dots), unless we impose unreasonable low temperatures ($T < 10^3$ K). We can slightly refine this calculation by considering that the Lyα forest probes densities higher than the mean, and thus we expect $\lambda_J^{(0)}$ to be lower. This is true given that the temperature is expected to scale with density as $T \propto \Delta^{\gamma-1}$, with $\gamma < 2$. The typical density of the Lyα forest as a function of redshift $\bar{\Delta}(z)$ and the temperature at such density $T_{\bar{\Delta}}(z)$ have been estimated recently using the "curvature" statistic by Boera et al. [2014] and Becker et al. [2011]. We use the two sets of values they obtain (labeled as $\bar{\Delta}_a, T_{\bar{\Delta},a}$ and $\bar{\Delta}_b, T_{\bar{\Delta},b}$, respectively) to produce the blue and the green curves. Despite of the slight improvement, the predictions from the classic Jeans formula are still overestimating the value of the filtering scale by about a factor of three.

The fact that the filtering scale should be smaller then the instantaneous Jeans scale has already been shown by Gnedin & Hui [1998] in the context of linear theory (see also chapter 1). The correction is typically a number between 1.5 and 3, but the precise value depends on the past thermal history. Moreover, linear theory is not reliable to model the Lyα forest, and hydrodynamic simulations are required to understand the precise relation between filtering scale and temperature evolution.

The solid lines in figure 6.4 represent the filtering scale fitted in the simulations at different redshifts and rescaled according to the procedure described in § 6.3. The NHM and N0.5HM simulations predict a higher Jeans scale then what we observe, and we need to assume a photoheating rate 10 times smaller than the standard Haardt and Madau model if we want to fit the measurement. Unfortunately, this model has a temperature at the mean density of $T_0 = 2505$ K, almost an order of magnitude than all the current estimates (see chapter 1). However, we stress again that this comparison relies on tuning the ratio $\lambda_{J,\mathrm{pairs}}/\lambda_{J,\mathrm{fit}}1.4$ between the the fitted cutoff of the real-flux power spectrum of simulation and the value that would be defined using the phase difference distributions calibrated on the set of DM-models. This correction has been tuned on the z=3 snapshot of NHM and it may not apply at other redshifts and in different runs. A rigorous comparison would require measuring the Jeans scale of the other simulations by applying the phase method to mock samples of pairs at all redshifts, a test we defer to future work.

Understanding the origin of such a small filtering scale will require more theoretical exploration, however we can speculate on possible explanations.

The Jeans filtering scale is sensitive to the past thermal history, and responds to temperature changes on timescales of the Hubble time. If at higher redshifts the temperature was lower than what assumed in the simulations, λ_J would retain memory of this cold stage and remain at lower values. Whether this could quantitatively explain the discrepancy shown in figure 6.4 should be verified in the future by means of hydrodynamic simulation with different thermal history. An alternative explanation would be some cosmological factor that enhanced the power of perturbations at small scales. It has been argued that primordial magnetic fields could have produced a similar effect [Subramanian & Barrow, 1998, Wasserman, 1978].

From the observational point of view, it is possible that there are systematic in the data which we are not correctly accounting for that decreases the coherence, inducing an underestimation of the Jeans scale. The most obvious candidates are metals and LLSs. Although their contribution to the Lyα forest absorption is modest, we have not explicitly demonstrated that their effect on phases is negligible.

6.5 Future Work

The work conducted so far has opened several important questions and posed few puzzles to our understanding of the IGM. The most relevant to solve is the explanation of the tiny Jeans scale that quasar pairs seem to indicate. The answer should be sought on a double track: theoretically and observationally. From the theoretical point of view, it is a priority to reach a consistent definition of the Jeans filtering scale of the IGM that could be used in hydrodynamic simulation and that could be related to the phase-difference method. In doing this, we might explore further the reliability of the approximated DM-based models on which the measurement is calibrated and characterize precisely the differences with hydrodynamic simulations. We will also need to understand to which extent the filtering scale is sensitive to the thermal history of higher redshifts, and whether a filtering scale consistent with our measurement can be achieved without making the IGM exceedingly cold.

From the observational perspective, it will be important to assess quantitatively the impact of systematic that we have not modeled properly, such as metals and LLSs. It is also crucial to understand the origin of the degeneracy between λ_J and γ that holds at $z = 2$ and $z = 2.4$, and whether this degeneracy can be broken by crossing the phase difference statistic with line-of-sight statistics. At the same time, we will collect new data in order to extend the analysis to higher redshifts $z > 3.3$ and to improve the statistical significance of the results presented here.

Chapter 7

Concluding Remarks

In this thesis I presented the first measurement of the Jeans filtering scale of the intergalactic medium. This filtering scale corresponds to the coherence length of the baryons set by the interplay between gravity and pressure across the history of the universe. It has fundamental cosmological implications: it provides a thermal record of heat injected by ultraviolet photons during cosmic reionization events, determines the clumpiness of the IGM, a critical ingredient in reionization models, and sets the minimum mass of galaxies to gravitationally collapse from the IGM. We elaborated a novel method to directly estimate the Jeans scale from the transverse coherence of Lyα absorption in quasar pair spectra. Our technique is based on the probability distribution function (PDF) of phase angle differences of homologous longitudinal Fourier modes in the spectra of the pair.

To study the efficacy of this new method, we combined a semi-analytical model of the Lyα forest with a dark matter only simulation, to generate a grid of 500 thermal models, where the temperature at mean density T_0, slope of the temperature-density relation γ, and the Jeans scale λ_J were varied. A Bayesian formalism is introduced based on the phase angle PDF, and MCMC techniques are used to conduct a full parameter study, allowing us to characterize the precision of a Jeans scale measurement, explore degeneracies with the other thermal parameters, and compare parameter constraints with those obtained from other statistics such as the longitudinal power and the cross-power spectrum.

The primary conclusions of this study are:

- The longitudinal power is highly degenerate with respect to the thermal parameters T_0, γ and λ_J, which arises because thermal broadening smooths the IGM along the line-of-sight (1D) at a comparable scale as the Jeans pressure smoothing (3D). It is

extremely challenging to disentangle this confluence of 1D and 3D smoothing with longitudinal observations alone. Similar degeneracies are likely to exist in other previously considered statistics sensitive to small-scale power such as the wavelet decomposition, the curvature, the b-parameter distribution, and the flux PDF. Hence it may be necessary to reassess the reliability and statistical significance of previous measurements of T_0 and γ.

- The cross-power measured from close quasar pairs is sensitive to the 3D Jeans smoothing, and can break degeneracies with the unknown Jeans scale. However, it is not the optimal statistic, because it mixes 1D information in the moduli of longitudinal Fourier modes, with the 3D information encoded in their phase differences. We show that by focusing on the phase differences alone, via the full PDF of phase angles, one is much more sensitive to 3D power and the Jeans smoothing.

- Based on a simple heuristic geometric argument, we derived an analytical form for the phase angle PDF. A single parameter family of wrapped-Cauchy distributions provides a good fit to the phase differences in our simulated spectra for any k, $r_\perp$, the full range of T_0,γ and λ_J.

- Our phase angle PDFs indicate that phase differences between large-scale longitudinal modes with small wavenumbers $k \ll 1/\lambda_J$, are nevertheless very sensitive to the Jeans scale. We present a simple analytical argument showing that this sensitivity results from the geometry of observing a 3D field along 1D skewers: low-k cross-power across correlated 1D skewers is actually dominated by high-k 3D modes up to a scale set by the pair separation $k_\perp \sim 1/r_\perp$.

- The phase angle PDF is essentially independent of the temperature-density relation parameters T_0 and γ. This results because 1) the non-linear FGPA transformation is only weakly dependent on temperature 2) phase angles of longitudinal modes are invariant to the symmetric thermal broadening convolution.

- Our full Bayesian MCMC parameter analysis indicates that a realistic sample of only 20 close quasar pair spectra observed at modest signal-to-noise ratio S/N $\simeq$ 10 and resolution of FWHM=30 km/s, can pinpoint the Jeans scale to $\simeq$ 5% precision, fully independent of the amplitude T_0 and slope γ of the temperature-density relation. The freedom from degeneracies with T_0 and γ is a direct consequence of the near independence of the phase angle PDF of these parameters.

- Our new estimator for the Jeans scale is unbiased and insensitive to a battery of systematics that typically plague Lyα forest measurements, such as continuum

fitting errors, imprecise knowledge of the noise level and/or spectral resolution, and metal-line absorption.

Motivated by these results, we applied the phase-difference technique to the existent sample of close quasar pairs. Adapting the method to real spectra requires significant modifications, among which:

- Calculation of phase differences on irregular grids by means of least-square spectral analysis. We have checked that alternative approximate methods (interpolation on regular grids) lead to the same results, guaranteeing their reliability
- Careful modeling of noise and resolution, and removal of the most evident contaminants such as broad absorption lines systems and damped Lyα absorbers.
- Calibration of the dynamic range in Fourier space according to the noise and resolution property of each spectrum, in order to exclude the noisiest mode at high k.

We performed the measurement in three different redshift bins, defined by the intervals $[1.8, 2.2], [2.2, 2.6]$ and $[2.7, 3.3]$. The phase difference analysis gives $\lambda_J = 66 \pm 20$ kpc at $z = 3$, $\lambda_J = 52 \pm 17$ kpc at $z = 3$ and $\lambda_J = 64 \pm 17$ kpc at $z = 3$. The current accuracy is at the level of 30% at all redshifts. At $z = 2$ and $z = 2.4$ the precision is decreased by a slight degeneracy of λ_J with γ. Interestingly, the direction of this degeneracy appears to be almost perpendicular to the same degeneracy expected from the line-of-sight power spectrum, a promising result in the perspective of crossing our constraints with other Lyα forest statistics.

We have tested that the results are stable with respect to the estimation of noise and resolution with a tolerance of about 10%. Most important, in the light of our results, the Jeans scale is hardly *under*estimated due to wrong noise/resolution assumptions. We also verified that phase differences are not sensitive to uncertainties on continuum placement.

In order for the parameter study presented here, with a large grid (500) of thermal models, to be computationally feasible, we had to rely on a simplified model of the IGM, based on a dark-matter only simulation and simple thermal scaling relations. In particular, the impact of Jeans pressure smoothing on the distribution of baryons is approximated by smoothing the dark-matter particle distribution with a Gaussian-like kernel, and we allowed the three thermal parameters T_0, γ, and λ_J to vary completely independently. Although the Gaussian filtering approximation is valid in linear theory [Gnedin et al., 2003], the Jeans scale is highly nonlinear at $z \simeq 3$, hence a precise

description of how pressure smoothing alters the 3D power spectrum of the baryons requires full hydrodynamical simulations. Furthermore, the three thermal parameters we consider are clearly implicitly correlated by the underlying thermal history of the Universe. Indeed, a full treatment of the impact of impulsive reionization heating on the thermal evolution of the IGM and the concomitant hydrodynamic response of the baryons, probably requires coupled radiative transfer hydrodynamical simulations.

Our approximate IGM model is thus justified by the complexity and computational cost of fully modeling the Jeans smoothing problem, and despite its simplicity, it provides a good fit to current measurements of the longitudinal power (see Figure 2.2). Most importantly, by analyzing the spatial structure of the *real-space flux* in hydrodynamic simulation, we proofed that calibrating phases on our simplified models is sufficient to locate the cutoff in the power spectrum of the low-density IGM.

A preliminary comparison with the expectations from hydrodynamic simulations indicates that the filtering scale we measured is too small to be explained with the standard assumptions on the thermal history and on the small-scale physics of the IGM. This discrepancy motivates further work to understand the theoretical implications of our findings, and demands a careful search for further systematics that could affect the phase difference statistic.

Appendix A

Resolving the Jeans Scale with Dark-Matter Simulations

The Lyα forest probes the structure of the very low density regions of the IGM, setting strict requirements on the resolution of our dark-matter only simulation. In particular, because our simulation is discrete in mass, each dark-matter particle represents a fixed amount of gas distributed according to the gravitational softening length and the size of the smoothing kernel that we use to represent Jeans smoothing (eqn. 2.3). At very low densities, it is possible that a very large region is described by a single particle, and that most of this void region is left empty. This undesirable situation occurs when the mean inter-particle separation $\Delta l = L_{\rm box}/N_{\rm p}^{1/3}$, which defines the typical size of regions occupied by each particle, is much larger than the Jeans scale λ_J, which is the minimum scale we want to resolve. Under such circumstances the density profile of skewers through our simulation cube will have many pixels which are nearly empty, because they have few or no neighboring particles. This insufficient sampling of the volume due to large mean inter-particle separation will then manifest itself through the appearance of artifacts in the volume-weighted probability distribution function (PDF) of the density. On the other hand, if the inter-particle separation is sufficiently small, the density field will be sufficiently sampled, and further decreasing the inter-particle separation will not alter the density PDF. Therefore we can define our resolution criteria for the mean inter-particle separation to be smaller than some multiple of the Jeans scale $\Delta l < \alpha\lambda_J$, where the exact value of this coefficient α is determined by checking that convergence is achieved in the density PDF.

We estimate α by plotting the PDF of $\log(\Delta)$ from our IGM skewers for a set of simulations with varying mean inter-particle separation, where $\Delta = \rho/\bar{\rho} = 1 + \delta$ is the density in units of the mean. The employed simulations have mean inter-particle separations $\Delta l = \{86, 171, 653\}$ kpc, corresponding to box sizes $L_{\rm box} = \{100, 250, 720\}$ Mpc/h with $N_{\rm p} = \{1500^3, 2048^3, 1800^3\}$ particles, respectively. In Figure A.1 we check for convergence using three different values of λ_J. The results indicate that a safe criterion for resolving the jeans scale is $\Delta l < \lambda_J$ or $\alpha \simeq 1$. The simulation employed in this work

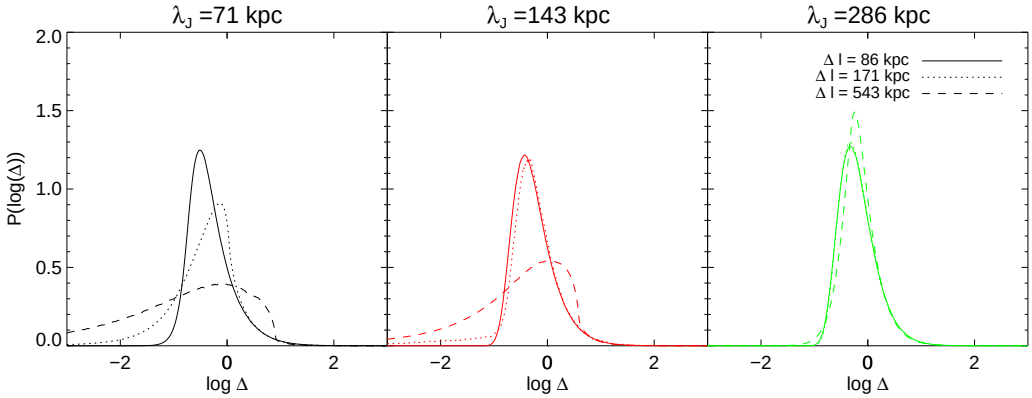

FIGURE A.1: The probability distributions of the relative baryonic density $\Delta = \rho/\bar{\rho}$ in our simulations. Each panel represent a different filtering scale λ_J, which was used to smooth the dark matter density for the same three simulations, which have different mean inter-particle separations Δl. When Δl is too large relative to λ_J the IGM density is poorly resolved at low densities, and the PDF is not converged. Empirically, we find that a safe criterion for convergence is $\Delta l < \lambda_J$, which allow us to resolve Jeans scales down to 50 kpc with our $L_{\rm box} = 50\,h^{-1}\,$Mpc and $N_{\rm p} = 1500^3$ simulation.

has $L_{\rm box} = 50\,h^{-1}\,$Mpc and $N_{\rm p} = 1500^3$ particles, or a mean inter-particle separation of $\Delta l = 48\,$kpc. This simulation thus allows us to study pressure smoothing down to a Jeans scale as small as $\simeq 50\,$kpc. Note however that the results of this paper rely on our estimation of the Jeans scale from various Lyα forest statistics around the fiducial value of $\lambda_J = 110\,$kpc, so we are confident that the Jeans scale is resolved in our simulations and that our results are not impacted by resolution effects.

Appendix B

Determining the Concentration Parameter ζ of the Wrapped-Cauchy Distribution

For a given sample of phases $\{\theta\}$ we employ a maximum-likelihood algorithm to determine the best-fit concentration parameter ζ, which uniquely specifies a wrapped-Cauchy distribution. This procedure is described in detail in Jammalamadaka & Sengupta [2001]. Briefly, we first reparametrize the wrapped-Cauchy distribution (eqn. 3.6) by writing $\nu = 2\zeta/(1+\zeta^2)$, which gives

$$P(\theta) \propto \frac{1}{1-\nu\cos(\theta)} \equiv w(\theta|\nu). \tag{B.1}$$

Following the standard recipe of maximizing the logarithm of the likelihood with respect to the desired parameter, we sum the logarithms of the probability of all angles and impose the condition that its derivative with respect to ν is zero, resulting in the equation

$$\sum_{i=1}^{n} w(\theta_i|\nu)[\cos(\theta_i) - \nu] = 0, \tag{B.2}$$

which can be solved iteratively. The concentration parameter is then easily determined by inverting the above relation to get $\zeta = (1-\sqrt{1-\nu^2})/\nu$. This procedure is repeated for each distinct population of phases, parametrized by transverse separation $r_\perp$ and k-mode, $\theta(r_\perp, k)$, and for each model in the thermal parameter grid (T_0, γ, λ_J) that we consider.

Appendix C

Phase Noise Calculation

In this appendix we show the derivation of formula 4.17. This formula expresses the probability distribution of the phase variation due to noise on a Fourier mode with coefficient F_0. We assume for simplicity that F_0 is real, without loss of generality since the calculation that follows is invariant under rotation of the complex plane. The noise in Fourier space F_N can be regarded as a Gaussian 2d stochastic variable with variance σ^2. To simplify the calculation, we renormalize all the moduli (the distances in the complex plane) such that $|F_0| = 1$. This operation is allowed because it does not change angles. With these assumptions $F_0 = 1$ is real and has unitary modulus. The rescaled variance of the noise will be $\eta^2 = \sigma^2/|F_0|^2$. We now want to obtain the probability distribution function in the complex plan of the variable $F' = F_0 + F_N$, which represent the Fourier coefficient after adding noise. Writing $F' = x + iy$ in the Cartesian representation and with the aforementioned assumption of the noise, we get

$$p_N(x+iy) = \frac{1}{2\pi\eta^2} exp\left[-\frac{(1-x)^2+y^2}{2\eta^2}\right]. \tag{C.1}$$

We now transform this probability function in polar coordinates, that gives

$$p_N(r,\phi) = \frac{r}{2\pi\eta^2}\exp\left[-\frac{r^2+1-2r\cos\phi}{2\eta^2}\right]. \tag{C.2}$$

where ϕ is exactly the phase variation with respect to the noiseless mode (see also figure C.1). If we want to calculate the distribution for ϕ we need to integrate in r to marginalize it out:

$$p_N(\phi) = \int_0^{+\infty} p_N(r,\phi)dr. \tag{C.3}$$

To solve the integral, we first rewrite the exponent using the identity $r^2+1-2r\cos\phi = (r-\cos\phi)^2+\sin^2\phi$, which leads to the expression

$$p_N(\phi) = \frac{e^{-\sin^2\phi/2\eta^2}}{2\pi\eta^2}\int_0^{+\infty} re^{-(r-\cos\phi)^2/2\eta^2}dr. \tag{C.4}$$

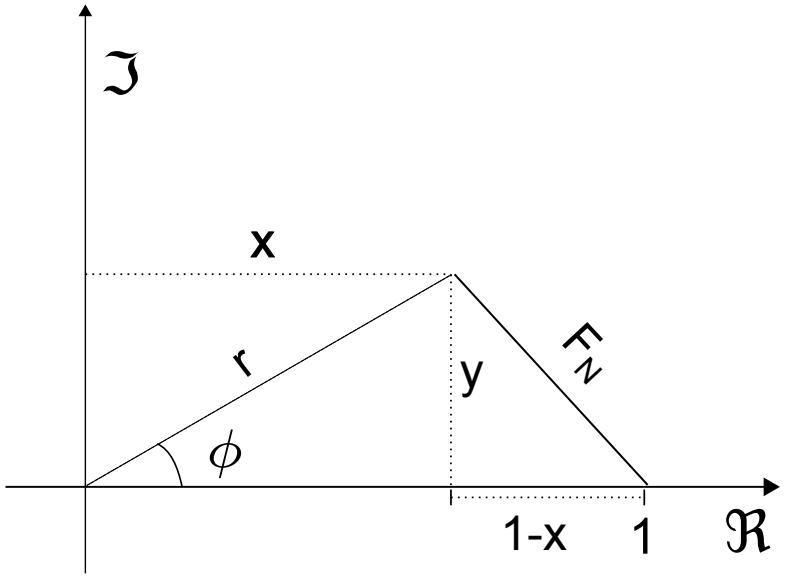

FIGURE C.1: Schematic representation in the complex plane of the relation of the noise displacements and the variation ϕ induced on the phase of the noiseless Fourier coefficient F_0. Since we are interested in the angular information, we scaled all the moduli and we rotate the plane such that $F_0 = 1$. x and y are the coordinates in the complex plane of the noisy mode $F' = F_0 + F_N$, while r and ϕ are the modulus and the phase of the polar representation of F'.

We then apply the change of variable $t = r - \cos\phi$ to the integral is split in two parts:

$$p_N(\phi) = \frac{e^{-\sin^2\phi/2\eta^2}}{2\pi\eta^2}\left[\int_{-\cos\phi}^{+\infty} te^{-t^2/2\eta^2}\,dt + \cos\phi\int_{-\cos\phi}^{+\infty} e^{-t^2/2\eta^2}\,dt\right]. \tag{C.5}$$

The first of the two integral can be easily solved with standard techniques, while the second can be recognized to be the complementary error function of the Gaussian distribution. After few lines of calculation, we finally obtain formula 4.17:

$$p_N(\phi) = \frac{e^{-1/2\eta^2}}{2\pi} + \frac{\cos\phi}{\sqrt{8\pi}}e^{-\sin^2\phi/2\eta^2}\operatorname{erfc}\left(-\frac{\cos\phi}{\sqrt{2\eta^2}}\right). \tag{C.6}$$

Bibliography

Abazajian, K. N., Adelman-McCarthy, J. K., Agüeros, M. A., Allam, S. S., Allende Prieto, C., An, D., Anderson, K. S. J., Anderson, S. F., Annis, J., Bahcall, N. A., & et al. 2009, ApJS, 182, 543

Abel, T., & Haehnelt, M. G. 1999, ApJ, 520, L13

Ade, P., et al. 2014, Astron.Astrophys.

Ahn, C. P., Alexandroff, R., Allende Prieto, C., Anderson, S. F., Anderton, T., Andrews, B. H., Aubourg, E., Bailey, S., Balbinot, E., Barnes, R., & et al. 2012, ApJS, 203, 21

Almgren, A. S., Bell, J. B., Lijewski, M. J., Lukić, Z., & Van Andel, E. 2013, ApJ, 765, 39

Barkana, R., & Loeb, A. 1999, ApJ, 523, 54

—. 2001, Phys. Rep., 349, 125

Becker, G. D., Bolton, J. S., Haehnelt, M. G., & Sargent, W. L. W. 2011, MNRAS, 410, 1096

Benson, A. J., Frenk, C. S., Lacey, C. G., Baugh, C. M., & Cole, S. 2002a, MNRAS, 333, 177

Benson, A. J., Lacey, C. G., Baugh, C. M., Cole, S., & Frenk, C. S. 2002b, MNRAS, 333, 156

Benson, A. J., & Madau, P. 2003, MNRAS, 344, 835

Boera, E., Murphy, M. T., Becker, G. D., & Bolton, J. S. 2014, MNRAS, 441, 1916

Bolton, J., Meiksin, A., & White, M. 2004, MNRAS, 348, L43

Bolton, J. S., Viel, M., Kim, T., Haehnelt, M. G., & Carswell, R. F. 2008, MNRAS, 386, 1131

Bovy, J., Hennawi, J. F., Hogg, D. W., Myers, A. D., Kirkpatrick, J. A., Schlegel, D. J., Ross, N. P., Sheldon, E. S., McGreer, I. D., Schneider, D. P., & Weaver, B. A. 2011, ApJ, 729, 141

Bovy, J., Myers, A. D., Hennawi, J. F., Hogg, D. W., McMahon, R. G., Schiminovich, D., Sheldon, E. S., Brinkmann, J., Schneider, D. P., & Weaver, B. A. 2012, ApJ, 749, 41

Broderick, A. E., Chang, P., & Pfrommer, C. 2012, ApJ, 752, 22

Bryan, G. L., & Machacek, M. E. 2000, ApJ, 534, 57

Bullock, J. S., Kravtsov, A. V., & Weinberg, D. H. 2000, ApJ, 539, 517

Calura, F., Tescari, E., D'Odorico, V., Viel, M., Cristiani, S., Kim, T.-S., & Bolton, J. S. 2012, MNRAS, 422, 3019

Cen, R., & Bryan, G. L. 2001, ApJ, 546, L81

Cen, R., McDonald, P., Trac, H., & Loeb, A. 2009, ArXiv e-prints

Cen, R., Miralda-Escudé, J., Ostriker, J. P., & Rauch, M. 1994, ApJ, 437, L9

Cen, R., & Ostriker, J. P. 1999, ApJ, 514, 1

—. 2006, ApJ, 650, 560

Chang, P., Broderick, A. E., & Pfrommer, C. 2012, ApJ, 752, 23

Chiang, L.-Y., Coles, P., & Naselsky, P. 2002, MNRAS, 337, 488

Ciardi, B., & Ferrara, A. 2005, Space Sci. Rev., 116, 625

Ciardi, B., & Salvaterra, R. 2007, MNRAS, 381, 1137

Coles, P. 2009, in Lecture Notes in Physics, Berlin Springer Verlag, Vol. 665, Data Analysis in Cosmology, ed. V. J. Martínez, E. Saar, E. Martínez-González, & M.-J. Pons-Bordería, 493–522

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

Croft, R. A. C., Weinberg, D. H., Katz, N., & Hernquist, L. 1997, ApJ, 488, 532

—. 1998, ApJ, 495, 44

Croom, S. M., Smith, R. J., Boyle, B. J., Shanks, T., Miller, L., Outram, P. J., & Loaring, N. S. 2004, MNRAS, 349, 1397

Davé, R., Cen, R., Ostriker, J. P., Bryan, G. L., Hernquist, L., Katz, N., Weinberg, D. H., Norman, M. L., & O'Shea, B. 2001, ApJ, 552, 473

Davé, R., Hernquist, L., Katz, N., & Weinberg, D. H. 1999, ApJ, 511, 521

Dijkstra, M., Haiman, Z., Rees, M. J., & Weinberg, D. H. 2004, ApJ, 601, 666

D'Odorico, V., Viel, M., Saitta, F., Cristiani, S., Bianchi, S., Boyle, B., Lopez, S., Maza, J., & Outram, P. 2006, MNRAS, 372, 1333

Emberson, J. D., Thomas, R. M., & Alvarez, M. A. 2013, ApJ, 763, 146

Fan, X., Carilli, C. L., & Keating, B. 2006, ARA&A, 44, 415

Faucher-Giguere, C. ., Prochaska, J. X., Lidz, A., Hernquist, L., & Zaldarriaga, M. 2007, ArXiv e-prints, 709

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2012, ArXiv e-prints

Furlanetto, S. R., & Dixon, K. L. 2010, ApJ, 714, 355

Furlanetto, S. R., & Oh, S. P. 2009, ApJ, 701, 94

Garzilli, A., Bolton, J. S., Kim, T.-S., Leach, S., & Viel, M. 2012, MNRAS, 424, 1723

Gnedin, N. Y. 2000a, ApJ, 542, 535

—. 2000b, ApJ, 542, 535

Gnedin, N. Y., Baker, E. J., Bethell, T. J., Drosback, M. M., Harford, A. G., Hicks, A. K., Jensen, A. G., Keeney, B. A., Kelso, C. M., Neyrinck, M. C., Pollack, S. E., & van Vliet, T. P. 2003, ApJ, 583, 525

Gnedin, N. Y., & Hui, L. 1998, MNRAS, 296, 44

Gunn, J. E., & Peterson, B. A. 1965, ApJ, 142, 1633

Haardt, F., & Madau, P. 2012, ApJ, 746, 125

Habib, S., Heitmann, K., Higdon, D., Nakhleh, C., & Williams, B. 2007, Phys. Rev. D, 76, 083503

Haehnelt, M. G., & Steinmetz, M. 1998, MNRAS, 298, L21

Haiman, Z., Abel, T., & Madau, P. 2001, ApJ, 551, 599

Heitmann, K., Higdon, D., Nakhleh, C., & Habib, S. 2006, ApJ, 646, L1

Heitmann, K., Higdon, D., White, M., Habib, S., Williams, B. J., Lawrence, E., & Wagner, C. 2009, ApJ, 705, 156

Heitmann, K., Lukić, Z., Fasel, P., Habib, S., Warren, M. S., White, M., Ahrens, J., Ankeny, L., Armstrong, R., O'Shea, B., Ricker, P. M., Springel, V., Stadel, J., & Trac, H. 2008, Computational Science and Discovery, 1, 015003

Heitmann, K., White, M., Wagner, C., Habib, S., & Higdon, D. 2010, ApJ, 715, 104

Hennawi, J. F. 2004, PhD thesis, Ph.D dissertation, 2004. 232 pages; United States – New Jersey: Princeton University; 2004. Publication Number: AAT 3151085. DAI-B 65/10, p. 5189, Apr 2005

Hennawi, J. F., Myers, A. D., Shen, Y., Strauss, M. A., Djorgovski, S. G., Fan, X., Glikman, E., Mahabal, A., Martin, C. L., Richards, G. T., Schneider, D. P., & Shankar, F. 2009, ArXiv e-prints

Hennawi, J. F., & Prochaska, J. X. 2007, ApJ, 655, 735

—. 2008, ArXiv Astrophysics e-prints

Hennawi, J. F., Prochaska, J. X., Burles, S., Strauss, M. A., Richards, G. T., Schlegel, D. J., Fan, X., Schneider, D. P., Zakamska, N. L., Oguri, M., Gunn, J. E., Lupton, R. H., & Brinkmann, J. 2006a, ApJ, 651, 61

Hennawi, J. F., Strauss, M. A., Oguri, M., Inada, N., Richards, G. T., Pindor, B., Schneider, D. P., Becker, R. H., Gregg, M. D., Hall, P. B., Johnston, D. E., Fan, X., Burles, S., Schlegel, D. J., Gunn, J. E., Lupton, R. H., Bahcall, N. A., Brunner, R. J., & Brinkmann, J. 2006b, AJ, 131, 1

Hoeft, M., Yepes, G., Gottlöber, S., & Springel, V. 2006, MNRAS, 371, 401

Hui, L., & Gnedin, N. Y. 1997, MNRAS, 292, 27

Hui, L., & Haiman, Z. 2003, ApJ, 596, 9

Hui, L., Stebbins, A., & Burles, S. 1999, ApJ, 511, L5

Iliev, I. T., Mellema, G., Pen, U.-L., Merz, H., Shapiro, P. R., & Alvarez, M. A. 2006, MNRAS, 369, 1625

Iliev, I. T., Scannapieco, E., & Shapiro, P. R. 2005, ApJ, 624, 491

Inoue, A. K., & Kamaya, H. 2003, MNRAS, 341, L7

Jakobsen, P., Boksenberg, A., Deharveng, J. M., Greenfield, P., Jedrzejewski, R., & Paresce, F. 1994, Nature, 370, 35

Jammalamadaka, S. R., & Sengupta, A. 2001

Kim, T.-S., Bolton, J. S., Viel, M., Haehnelt, M. G., & Carswell, R. F. 2007, MNRAS, 382, 1657

Kim, T.-S., Viel, M., Haehnelt, M. G., Carswell, R. F., & Cristiani, S. 2004, MNRAS, 347, 355

Kollmeier, J. A., Miralda-Escudé, J., Cen, R., & Ostriker, J. P. 2006, ApJ, 638, 52

Kuhlen, M., & Madau, P. 2005, MNRAS, 363, 1069

Kulkarni, G., & Choudhury, T. R. 2011, MNRAS, 412, 2781

Kwan, J., Bhattacharya, S., Heitmann, K., & Habib, S. 2012, ArXiv e-prints

Lawrence, E., Heitmann, K., White, M., Higdon, D., Wagner, C., Habib, S., & Williams, B. 2010, ApJ, 713, 1322

Lee, K.-G. 2012, ApJ, 753, 136

Lee, K.-G., Bailey, S., Bartsch, L. E., Carithers, W., Dawson, K. S., Kirkby, D., Lundgren, B., Margala, D., Palanque-Delabrouille, N., Pieri, M. M., Schlegel, D. J., Weinberg, D. H., Yèche, C., Aubourg, E., Bautista, J., Bizyaev, D., Blomqvist, M., Bolton, A. S., Borde, A., Brewington, H., Busca, N. G., Croft, R. A. C., Delubac, T., Ebelke, G., Eisenstein, D. J., Font-Ribera, A., Ge, J., Hamilton, J.-C., Hennawi, J. F., Ho, S., Honscheid, K., Le Goff, J.-M., Malanushenko, E., Malanushenko, V., Miralda-Escudé, J., Myers, A. D., Noterdaeme, P., Oravetz, D., Pan, K., Pâris, I., Petitjean, P., Rich, J., Rollinde, E., Ross, N. P., Rossi, G., Schneider, D. P., Simmons, A., Snedden, S., Slosar, A., Spergel, D. N., Suzuki, N., Viel, M., & Weaver, B. A. 2013, AJ, 145, 69

Lidz, A., Faucher-Giguere, C., Dall'Aglio, A., McQuinn, M., Fechner, C., Zaldarriaga, M., Hernquist, L., & Dutta, S. 2009, ArXiv e-prints

Lomb, N. R. 1976, Ap&SS, 39, 447

Lumsden, S. L., Heavens, A. F., & Peacock, J. A. 1989, MNRAS, 238, 293

Madau, P. 2000, in Royal Society of London Philosophical Transactions Series A, Vol. 358, Astronomy, physics and chemistry of H_3^+, 2021

Madau, P., & Efstathiou, G. 1999, ApJ, 517, L9

Madau, P., Ferrara, A., & Rees, M. J. 2001, ApJ, 555, 92

Madau, P., Haardt, F., & Rees, M. J. 1999, ApJ, 514, 648

Madau, P., & Meiksin, A. 1994, ApJ, 433, L53

Mathias, A., Grond, F., Guardans, R., Seese, D., Canela, M., & Diebner, H. H. 2004, Journal of Statistical Software, 11, 1

McDonald, P., Miralda-Escudé, J., Rauch, M., Sargent, W. L. W., Barlow, T. A., & Cen, R. 2001, ApJ, 562, 52

McDonald, P., Miralda-Escudé, J., Rauch, M., Sargent, W. L. W., Barlow, T. A., Cen, R., & Ostriker, J. P. 2000, ApJ, 543, 1

McDonald, P., Seljak, U., Burles, S., Schlegel, D. J., Weinberg, D. H., Cen, R., Shih, D., Schaye, J., Schneider, D. P., Bahcall, N. A., Briggs, J. W., Brinkmann, J., Brunner, R. J., Fukugita, M., Gunn, J. E., Ivezić, v., Kent, S., Lupton, R. H., & Vanden Berk, D. E. 2006, ApJS, 163, 80

McDonald, P., Seljak, U., Cen, R., Bode, P., & Ostriker, J. P. 2005, MNRAS, 360, 1471

McGill, C. 1990, MNRAS, 242, 544

McQuinn, M., Lidz, A., Zaldarriaga, M., Hernquist, L., Hopkins, P. F., Dutta, S., & Faucher-Giguère, C.-A. 2009, ApJ, 694, 842

McQuinn, M., Oh, S. P., & Faucher-Giguère, C.-A. 2011, ApJ, 743, 82

Meiksin, A., & Tittley, E. R. 2012, MNRAS, 423, 7

Meiksin, A., & White, M. 2001, MNRAS, 324, 141

Meiksin, A. A. 2009, Reviews of Modern Physics, 81, 1405

Miralda-Escudé, J., Cen, R., Ostriker, J. P., & Rauch, M. 1996, ApJ, 471, 582

Miralda-Escudé, J., Haehnelt, M., & Rees, M. J. 2000, ApJ, 530, 1

Miralda-Escudé, J., & Rees, M. J. 1994, MNRAS, 266, 343

Myers, A. D., Richards, G. T., Brunner, R. J., Schneider, D. P., Strand, N. E., Hall, P. B., Blomquist, J. A., & York, D. G. 2008, ApJ, 678, 635

Nath, B. B., & Biermann, P. L. 1993, MNRAS, 265, 241

Nath, B. B., Sethi, S. K., & Shchekinov, Y. 1999, MNRAS, 303, 1

Okamoto, T., Gao, L., & Theuns, T. 2008, MNRAS, 390, 920

Parsons, A. R., Liu, A., Aguirre, J. E., Ali, Z. S., Bradley, R. F., Carilli, C. L., DeBoer, D. R., Dexter, M. R., Gugliucci, N. E., Jacobs, D. C., Klima, P., MacMahon, D. H. E., Manley, J. R., Moore, D. F., Pober, J. C., Stefan, I. I., & Walbrugh, W. P. 2013, ArXiv e-prints

Pawlik, A. H., Schaye, J., & van Scherpenzeel, E. 2009, MNRAS, 394, 1812

Peacock, J. A. 1999, Cosmological physics; rev. version (Cambridge: Cambridge Univ.)

Peeples, M. S., Weinberg, D. H., Davé, R., Fardal, M. A., & Katz, N. 2009a, ArXiv e-prints

—. 2009b, ArXiv e-prints

Penrose, R. 1955, Mathematical Proceedings of the Cambridge Philosophical Society, 51, 406

Petry, C. E., Impey, C. D., & Foltz, C. B. 1998, ApJ, 494, 60

Pfrommer, C., Chang, P., & Broderick, A. E. 2012, ApJ, 752, 24

Prochaska, J. X., & Hennawi, J. F. 2009, ApJ, 690, 1558

Prochaska, J. X., Hennawi, J. F., Lee, K.-G., Cantalupo, S., Bovy, J., Djorgovski, S. G., Ellison, S. L., Lau, M. W., Martin, C. L., Myers, A., Rubin, K. H. R., & Simcoe, R. A. 2013, ApJ, 776, 136

Prochaska, J. X., Hennawi, J. F., & Simcoe, R. A. 2012, ArXiv e-prints

Prochaska, J. X., O'Meara, J. M., & Worseck, G. 2010, ApJ, 718, 392

Puchwein, E., Pfrommer, C., Springel, V., Broderick, A. E., & Chang, P. 2012, MNRAS, 423, 149

Rauch, M. 1998, ARA&A, 36, 267

Rauch, M., Sargent, W. L. W., Barlow, T. A., & Carswell, R. F. 2001, ApJ, 562, 76

Reimers, D., Kohler, S., Wisotzki, L., Groote, D., Rodriguez-Pascual, P., & Wamsteker, W. 1997, A&A, 327, 890

Richards, G. T., Nichol, R. C., Gray, A. G., Brunner, R. J., Lupton, R. H., Vanden Berk, D. E., Chong, S. S., Weinstein, M. A., Schneider, D. P., Anderson, S. F., Munn, J. A., Harris, H. C., Strauss, M. A., Fan, X., Gunn, J. E., Ivezić, v., York, D. G., Brinkmann, J., & Moore, A. W. 2004, ApJS, 155, 257

Ricotti, M., Gnedin, N. Y., & Shull, J. M. 2000, ApJ, 534, 41

Ricotti, M., Ostriker, J. P., & Gnedin, N. Y. 2005, MNRAS, 357, 207

Rorai, A., Hennawi, J. F., & White, M. 2013, ApJ, 775, 81

Rudie, G. C., Steidel, C. C., & Pettini, M. 2012, ApJ, 757, L30

Scannapieco, E., Ferrara, A., & Madau, P. 2002, ApJ, 574, 590

Scannapieco, E., & Oh, S. P. 2004, ApJ, 608, 62

Schaye, J., Theuns, T., Rauch, M., Efstathiou, G., & Sargent, W. L. W. 2000, MNRAS, 318, 817

Shen, Y., Hennawi, J. F., Shankar, F., Myers, A. D., Strauss, M. A., Djorgovski, S. G., Fan, X., Giocoli, C., Mahabal, A., Schneider, D. P., & Weinberg, D. H. 2010, ApJ, 719, 1693

Shull, J. M., France, K., Danforth, C. W., Smith, B., & Tumlinson, J. 2010, ApJ, 722, 1312

Smette, A., Robertson, J. G., Shaver, P. A., Reimers, D., Wisotzki, L., & Koehler, T. 1995, A&AS, 113, 199

Somerville, R. S. 2002, ApJ, 572, L23

Springel, V. 2005, MNRAS, 364, 1105

Subramanian, K., & Barrow, J. D. 1998, Phys. Rev. D, 58, 083502

Syphers, D., Anderson, S. F., Zheng, W., Meiksin, A., Schneider, D. P., & York, D. G. 2012, AJ, 143, 100

Tanaka, T., Perna, R., & Haiman, Z. 2012a, MNRAS, 425, 2974

—. 2012b, MNRAS, 425, 2974

Theuns, T., Mo, H. J., & Schaye, J. 2001, MNRAS, 321, 450

Theuns, T., Schaye, J., & Haehnelt, M. G. 2000, MNRAS, 315, 600

Theuns, T., Schaye, J., Zaroubi, S., Kim, T., Tzanavaris, P., & Carswell, B. 2002a, ApJ, 567, L103

Theuns, T., Zaroubi, S., Kim, T., Tzanavaris, P., & Carswell, R. F. 2002b, MNRAS, 332, 367

Tittley, E. R., & Meiksin, A. 2007a, MNRAS, 380, 1369

—. 2007b, MNRAS, 380, 1369

Viel, M., Bolton, J. S., & Haehnelt, M. G. 2009, MNRAS, 399, L39

Viel, M., Matarrese, S., Mo, H. J., Haehnelt, M. G., & Theuns, T. 2002, MNRAS, 329, 848

Voit, G. M. 1996, ApJ, 465, 548

Wasserman, I. 1978, ApJ, 224, 337

Watts, P., Coles, P., & Melott, A. 2003, ApJ, 589, L61

White, M. 2002, ApJS, 143, 241

Worseck, G., Prochaska, J. X., McQuinn, M., Dall'Aglio, A., Fechner, C., Hennawi, J. F., Reimers, D., Richter, P., & Wisotzki, L. 2011, ApJ, 733, L24

York, D. G., Adelman, J., Anderson, J. J. E., Anderson, S. F., Annis, J., Bahcall, N. A., Bakken, J. A., Barkhouser, R., Bastian, S., Berman, E., Boroski, W. N., Bracker, S., Briegel, C., Briggs, J. W., Brinkmann, J., Brunner, R., Burles, S., Carey, L., Carr, M. A., Castander, F. J., Chen, B., Colestock, P. L., Connolly, A. J., Crocker, J. H., Csabai, I., Czarapata, P. C., Davis, J. E., Doi, M., Dombeck, T., Eisenstein, D., Ellman, N., Elms, B. R., Evans, M. L., Fan, X., Federwitz, G. R., Fiscelli, L., Friedman, S., Frieman, J. A., Fukugita, M., Gillespie, B., Gunn, J. E., Gurbani, V. K., de Haas, E., Haldeman, M., Harris, F. H., Hayes, J., Heckman, T. M., Hennessy, G. S., Hindsley, R. B., Holm, S., Holmgren, D. J., Huang, C.-h., Hull, C., Husby, D., Ichikawa, S.-I., Ichikawa, T., Ivezić, v., Kent, S., Kim, R. S. J., Kinney, E., Klaene, M., Kleinman, A. N., Kleinman, S., Knapp, G. R., Korienek, J., Kron, R. G., Kunszt, P. Z., Lamb, D. Q., Lee, B., Leger, R. F., Limmongkol, S., Lindenmeyer, C., Long, D. C., Loomis, C., Loveday, J., Lucinio, R., Lupton, R. H., MacKinnon, B., Mannery, E. J., Mantsch, P. M., Margon, B., McGehee, P., McKay, T. A., Meiksin, A., Merelli, A., Monet, D. G., Munn, J. A., Narayanan, V. K., Nash, T., Neilsen, E., Neswold, R., Newberg, H. J., Nichol, R. C., Nicinski, T., Nonino, M., Okada, N., Okamura, S., Ostriker, J. P., Owen, R., Pauls, A. G., Peoples, J., Peterson, R. L., Petravick, D., Pier, J. R., Pope, A., Pordes, R., Prosapio, A., Rechenmacher, R., Quinn, T. R., Richards, G. T., Richmond, M. W., Rivetta, C. H., Rockosi, C. M., Ruthmansdorfer, K., Sandford, D., Schlegel, D. J., Schneider, D. P., Sekiguchi, M., Sergey, G., Shimasaku, K., Siegmund, W. A., Smee, S., Smith, J. A., Snedden, S., Stone, R., Stoughton, C., Strauss, M. A., Stubbs, C., SubbaRao, M., Szalay, A. S., Szapudi, I., Szokoly, G. P., Thakar, A. R., Tremonti, C., Tucker, D. L., Uomoto, A., Vanden Berk, D., Vogeley, M. S., Waddell, P., Wang, S.-i., Watanabe, M., Weinberg, D. H., Yanny, B., & Yasuda, N. 2000, AJ, 120, 1579

Young, P., Sargent, W. L. W., Oke, J. B., & Boksenberg, A. 1981, ApJ, 249, 415

Zaldarriaga, M. 2002, The Astrophysical Journal, 564, 153

Zaldarriaga, M., Hui, L., & Tegmark, M. 2001, ApJ, 557, 519

Zaroubi, S. 2013, in Astrophysics and Space Science Library, Vol. 396, Astrophysics and Space Science Library, ed. T. Wiklind, B. Mobasher, & V. Bromm, 45

Acknowledgements

I am deeply thankful to my supervisor, Joseph Hennawi, first of all for offering me the chance of working on a challenging and fascinating problem during these years in Heidelberg. His guidance, beside guaranteeing a sane advancement of my research, helped me in developing a rigorous and critic attitude toward my own ideas and a better organized and efficient work style. I particularly value the fact that his advice went always far beyond the pure scientific scope of the project, showing an authentic interest in my general formation as a researcher.

This project would not have been possible without the support of prof. Martin White from the University of Berkeley, who provided us with the Dark Matter simulations we needed, and constantly discussed with us about the direction and the goals of the projects.

I am grateful to all the current and past members of the ENIGMA group for the friendly and stimulating working environment, and for the useful discussions during all the stages of my PhD. I wish to thank in particular Jose Oñorbe and Girish Kulkarni for their crucial contribution to the final phase of my project and for comments and suggestions that significantly improved this manuscript. I am also in debt with Xavier Prochaska for patiently leading me in my first steps in the insidious world of continuum fitting. It was really valuable the help of my friends and colleagues Rahul, Sladjana and Christina in sorting out the last logistic and formal details during the hasty days before thesis submission.

I want to thank the MPIA and IMPRS staffs for the great opportunities they guaranteed me, not only on the academic side, and for making my stay in Heidelberg a totally enjoyable experience.

I finally express my gratitude to all those persons who have been far from me in these years for most of the time, but did not stop caring of me, despite of knowing me so well. Among these, I want in particular to thank my parents, Daniela and Ernesto, my sister Cecilia, and Claudia.