Precisely measuring the cosmic reionization history from IGM damping wings towards quasars

Joseph F. Hennawi [1,2]⋆ Timo Kist [1] Frederick B. Davies[3] and John Tamanas[4]

[1]*Leiden Observatory, Leiden University, PO Box 9513, NL-2300 RA Leiden, The Netherlands*
[2]*Department of Physics, University of California, Santa Barbara, CA 93106, USA*
[3]*Max-Planck-Institut für Astronomie, K'onigstuhl 17, D-69117 Heidelberg, Germany*
[4]*Department of Physics, University of California, Santa Cruz, CA 95064, USA*

ABSTRACT

We introduce a new approach for analysing the intergalactic medium (IGM) damping wings imprinted on the proximity zones of quasars in the epoch of reionization (EoR). Whereas past work has typically forgone the additional constraining power afforded by the blue side continuum ($\lambda \lesssim 1280$ Å) and/or opted not to model the large correlated IGM transmission fluctuations in the proximity zone ($\lambda \lesssim 1216$ Å), we construct a generative probabilistic model for the entire spectrum accounting for all sources of error – the stochasticity induced by patchy reionization, the impact of the quasar's ionizing radiation on the IGM, the unknown intrinsic spectrum of the quasar, and spectral noise. This principled Bayesian method allows us to marginalize out nuisance parameters associated with the quasar's radiation and its unknown intrinsic spectrum to precisely measure the IGM neutral fraction, $\langle x_{\mathrm{HI}} \rangle$. A key element of our analysis is the use of dimensionality reduction (DR) to describe the intrinsic quasar spectrum via a small number of nuisance parameters. Using a large sample of 15 559 SDSS/BOSS quasars at $z \gtrsim 2.15$ we trained and quantified the performance of six distinct DR methods, and find that a six parameter principal component analysis model (five coefficients plus a normalization) performs best, with complex machine-learning approaches providing no advantage. By conducting statistical inference on 100 realistic mock EoR quasar spectra, we demonstrate the reliability of the credibility contours that we obtain on $\langle x_{\mathrm{HI}} \rangle$ and the quasar lifetime, t_{Q}. The new method introduced here will transform IGM damping wings into a precision probe of reionization, on the same solid methodological and statistical footing as other precision cosmological measurements.

Key words: intergalactic medium – quasars: absorption lines – dark ages, reionization, first stars – cosmology: observations.

1 INTRODUCTION

About 380 000 yr after the big bang, primordial plasma recombined to form the first atoms, releasing the cosmic microwave background (CMB) and initiating the cosmic 'dark ages', which prevailed until radiation from stars and black holes in primeval galaxies reionized the Universe. Understanding how this epoch of reionization (EoR) emerged and the nature of the early sources that drove it are among the most important open questions in cosmology and key science drivers for numerous major observatories (*Planck*, LOFAR, SKA, HERA, Keck, VLT, *Hubble Space Telescope*, *Euclid*, and *JWST*).

Our understanding of the evolution of the average intergalactic medium (IGM) neutral fraction with cosmic time, $\langle x_{\mathrm{HI}}(z) \rangle$, currently rests upon two pillars. The first is the CMB electron scattering optical depth τ_e, which provides an integral constraint on $\langle x_{\mathrm{HI}}(z) \rangle$, but leaves its shape poorly determined. The second is the Gunn–Peterson (GP) Lyα opacity measured towards distant $z \gtrsim 6$ quasars, which only robustly constrains the end of reionization (e.g McGreer,

Mesinger & D'Odorico 2015; Jin et al. 2023), because the overly sensitive Lyα transition saturates for neutral fractions of $\langle x_{\mathrm{HI}} \rangle \gtrsim 10^{-4}$. The final *Planck* CMB constraints on reionization history, which incorporate the GP Lyα opacity constraint on the end of the reionization, indicate that the IGM was 50 per cent neutral at a redshift in the range $z_{\mathrm{reion}} \simeq 5.9 - 8.0$ (2σ; Planck Collaboration VI 2020), considerably lower than the $z_{\mathrm{reion}} \sim 17$ initially inferred by *WMAP* (*Wilkinson Microwave Anisotropy Probe*; Spergel et al. 2003), and pulling reionization into the realm of the highest-z quasars known.

Low-frequency radio observations of the 21-cm line have been touted as the premier probe of reionization and are steadily increasing in sensitivity (e.g. The HERA Collaboration 2022). They aim to detect the minuscule cosmic 21-cm background beneath foregrounds and instrumental systematics that are 10^5 times larger (Cheng et al. 2018), but have yet to provide quantitative constraints on $\langle x_{\mathrm{HI}}(z) \rangle$. Similarly, the so-called kinetic Sunyaev–Zeldovich (kSZ) effect, which must be disentangled from other small-scale secondary CMB anistropies as well as post-reionization kSZ contributions (e.g. Doré, Hennawi & Spergel 2004; Ferraro & Smith 2018), also holds promise, but current $\langle x_{\mathrm{HI}}(z) \rangle$ constraints are weak (Nikolić et al. 2023) and model-dependent (Zahn et al. 2012; George et al. 2015).

⋆ E-mail: joe@physics.ucsb.edu

In a neutral IGM, the GP optical depth is so large, $\tau_{Ly\alpha} \sim 10^5$, that the red Lorentzian wing of the Lyα absorption cross-section can imprint an observable IGM damping wing on the spectrum of a background astronomical source (Miralda-Escudé 1998). Multiple methods have been proposed to leverage this unique signature to obtain quantitative constraints on $\langle x_{HI}(z) \rangle$ using either high-z galaxies or quasars as the background sources. For quasars, one can either analyse the damped Lyα (DLA) absorption signature arising from the IGM in the immediate vicinity of the quasar, or in the foreground (closer to the observer) where it could be imprinted upon the lower-z fluctuating Lyα forest transmission (Malloy & Lidz 2015).[1] Whereas the former case is the subject of this paper, several recent studies investigate the latter in the vicinity of individual GP troughs embedded in the $5.5 \lesssim z \lesssim 6$ Lyα forest (Becker et al. 2024), or stacks thereof (Spina et al. 2024; Zhu et al. 2024) with the hope of constraining $\langle x_{HI}(z) \rangle$. In addition, IGM damping wings strongly suppress Lyα emission lines from EoR galaxies, and it has been argued that constraints on $\langle x_{HI}(z) \rangle$ can be obtained from the statistics of Lyα line strengths (Dijkstra, Mesinger & Wyithe 2011; Mesinger et al. 2015), motivating a large body of work exploring this technique (e.g. Mason et al. 2018, 2019; Hoag et al. 2019; Jung et al. 2020). Another approach recently enabled by the exquisite spectra of EoR galaxies provided by *JWST* is to try to measure the IGM damping wing signature from the continua of individual or stacked galaxy spectra (Umeda et al. 2024; Keating et al. 2024a).

However, Heintz et al. (2025, see also Heintz et al. 2024; Umeda et al. 2024; D'Eugenio et al. 2024; Chen et al. 2024) recently demonstrated that a significant fraction $\sim 60 - 70$ per cent of EoR galaxies with $5.5 \lesssim z \lesssim 13$ exhibit strong intrinsic DLA absorption with $N_{HI} \gtrsim 10^{21}$ cm^{-2}, far stronger than the corresponding intrinsic DLAs seen in star-forming galaxies at lower-z ($z \lesssim 4$), and importantly, than the expected damping wings from a neutral IGM (typical $N_{HI} \sim 10^{21}$ cm^{-2}). It is perhaps not surprising that this intrinsic absorption, which arises from the interstellar medium (ISM) or circumgalactic medium (CGM) of galaxies, should rapidly evolve as one approaches the EoR. Estimates of the ultraviolet background (UVB) from the statistics of Lyα forest transmission indicate rapid evolution towards $z \gtrsim 5.5$ (Davies et al. 2018a, 2024; Gaikwad et al. 2023), which is further supported by the similarly rapidly evolving mean-free path of ionizing photons (λ_{mfp}; Becker et al. 2021; Zhu et al. 2023), since the UVB is proportional to the mean-free path. CGM absorbers in ionization equilibrium with this evolving UVB are thus expected to be stronger and more abundant in the EoR (Bolton & Haehnelt 2013), which is empirically supported by both the strong increase in the abundance of low-ionization metal absorption lines at $z \gtrsim 6$ (Becker et al. 2019; Christensen et al. 2023) as well as the strong redshift evolution in the occurrence of DLAs in star-forming galaxies observed by Heintz et al. (2025). Regardless of the physical explanation, *the increased prevalence of strong intrinsic DLA absorption in EoR galaxies calls into question the entire enterprise of using galaxies, whether via the statistics of their Lyα lines or via damping wing absorption imprinted on their continua, as background sources to probe reionization.* No study to date has demonstrated that the intrinsic ISM/CGM absorption in galaxies can be disentangled from the damping wing absorption induced by the neutral IGM. Further obstacles arise from the poorly understood intrinsic nebular continuum shape of galaxy spectra near

the Lyα line (Raiter, Schaerer & Fosbury 2010; Byler et al. 2017), and from the challenge of modelling their Lyα emission lines, which is complicated by resonant scattering effects (Sadoun, Zheng & Miralda-Escudé 2017).

In contrast, quasars provide several advantages over galaxies as background sources for IGM damping wing measurements. First, quasars are far brighter, allowing one to obtain high signal-to-noise ratio (S/N $\gtrsim 10$) and high-resolution ($R \sim 3000 - 10000$) spectra with far less telescope time than required for galaxies. Second, analogous to the way O-stars transform their nearby ISM, the quasar's own ionizing radiation sources a giant Mpc-scale H II region (Cen & Haiman 2000) known as a proximity zone, manifest as enhanced Lyα transmission near the quasar itself (e.g. Fan et al. 2006; Eilers et al. 2017a) Although this reduces the strength of the IGM damping wing and adds a nuisance physical process to the modelling, it provides the great advantage that neutral gas in the ISM/CGM of the quasar's host galaxy is completely ionized away, eliminating systematics associated with poorly understood ISM/CGM absorption. This simplifies the modelling dramatically, as one only needs to treat 1D radiative transfer (RT) of the quasar radiation through the diffuse IGM, which is a well-posed problem that can be solved *ab initio* – IGM density fluctuations can be predicted from first principles and the IGM damping wing strength (being an integral constraint) is insensitive to the exact details of the reionization topology at fixed $\langle x_{HI} \rangle$ (Davies et al. 2018b; Chen 2024; Keating et al. 2024a). Furthermore, this approach is also largely insensitive[2] to the details of the quasar's radiative history (Davies et al. 2019, 2020), which can be encapsulated by a single nuisance parameter, t_Q, the quasar lifetime (assuming a 'light-bulb' light curve). The third advantage is that modelling the intrinsic quasar spectrum near Lyα is straightforward. Because there is very little evidence that quasar spectra evolve from $z \sim 2 - 7$ (Shen et al. 2019), large training sets of thousands of $z \sim 2 - 3$ quasar spectra from surveys like SDSS/DESI can be used to train empirical models for their intrinsic spectra (e.g. Greig et al. 2017a; Davies et al. 2018c).

The intrinsic quasar spectrum constitutes a nuisance stochastic process that must be marginalized out to obtain constraints on the astrophysical parameters $\langle x_{HI} \rangle$ and t_Q. Indeed, a critical aspect of measuring the IGM damping wing signature in EoR quasar spectra is obtaining a reliable estimate of and uncertainty for the intrinsic unabsorbed quasar spectrum in the region near the Lyα line. To date, the approach adopted in most past work is to use the spectrum redward of the absorbed Lyα region (e.g. $\lambda > 1280$ Å) to predict the blue part of the unabsorbed quasar spectrum in the vicinity of the

---

[1] In practice, the demarcation between these two regimes depends on the extent of the region that is overionized by the quasar, i.e. the size of its proximity zone.

[2] For a quasar shining into a completely neutral IGM, the size of its surrounding ionized region is primarily set by the total number of ionizing photons emitted over the age of the Universe, and is largely insensitive to its light curve due to the long recombination time-scale relative to the *Hubble* time. In contrast, for a quasar in a highly ionized IGM, the light curve can more strongly affect IGM transmission in the proximity zone. This sensitivity arises if the quasar's radiation varies on timescales comparable to the equilibrium time, $t_{eq} \equiv \Gamma_{HI}^{-1} \sim 3 \times 10^4$ yr, the time-scale on which the IGM responds to quasar radiation or recombines to its baseline neutral fraction in the absence of the quasar (Davies, Hennawi & Eilers 2020). Observations of quasar proximity zones are broadly consistent with a simple light-bulb light curve (Eilers et al. 2017a; Davies et al. 2020), with lognormally distributed lifetimes t_Q centred around $\sim 10^6$ yr and a standard deviation of ~ 1 dex (Eilers et al. 2020, 2021; Khrykin et al. 2021; Morey et al. 2021; Satyavolu et al. 2023; Ďurovčíková et al. 2024). Thus, even in an ionized IGM, current evidence suggests light-curve variability has limited impact on proximity zone transmission, though this remains an important topic for future study.

Lyα line (e.g. $\lambda < 1280$ Å), an idea first suggested by Suzuki et al. (2005) in the context of studies of the $z \lesssim 4$ Lyα forest (Pâris et al. 2011; Lee, Suzuki & Spergel 2012; Lee et al. 2015; Eilers et al. 2017a). Mortlock et al. (2011) were the first to apply this red-blue prediction approach to search for an IGM damping wing imprinted on a $z = 7.1$ quasar spectrum. This approach was elaborated upon by Greig et al. (2017a), who adopted a parametric model for the unabsorbed spectrum – a sum of a power law plus multiple Gaussian emission-line components – and used the statistical correlations between these parameters, inferred from a large training data set, to generate the red–blue predictions. Building on past Lyα forest work using principal component analysis (PCA; Pâris et al. 2011; Lee et al. 2012, 2015; Eilers et al. 2017a), Davies et al. (2018b) developed a PCA-based red–blue continuum prediction algorithm (see also Eilers, Hennawi & Lee 2017b), trained on a set of $> 10\,000$ quasar spectra at $z \gtrsim 2$ from the SDSS/BOSS (Baryon Oscillation Spectroscopic Survey) DR12 quasar catalogue (Pâris et al. 2017). They quantified the covariant uncertainties in the predicted continua, and found them to have a typical relative error of $\sim 6 - 12$ per cent at the rest-frame wavelengths most relevant for IGM damping wing absorption. These studies motivated the development of a plethora of red–blue continuum prediction methods for studying quasar proximity zones and IGM damping wings (Ďurovčíková et al. 2020; Fathivavsari 2020; Reiman 2020; Chen et al. 2022) as well as the Lyα forest (Bosman et al. 2021; Liu & Bordoloi 2021). Greig et al. (2024b) recently conducted a detailed comparison of the ~ 10 red–blue quasar continuum prediction pipelines in existence (see also Bosman et al. 2021), and found that they all yield roughly comparable precision of $\sim 10 - 20$ per cent at the relevant blue-side rest-frame wavelengths near the Lyα line.

The discovery of the first $z > 7$ quasar ULASJ1120+0641 (Mortlock et al. 2011) led to significant interest in using such sources to obtain quantitative constraints on reionization (Mortlock et al. 2011; Bolton et al. 2011; Keating et al. 2015; Greig et al. 2017b). Greig et al. (2017b) combined the Greig et al. (2017a) continuum prediction algorithm with seminumerical simulations of the reionization topology (Mesinger, Greig & Sobacchi 2016) to model the distribution of IGM damping wing strengths as a function of $\langle x_{\mathrm{HI}}(z) \rangle$, but their analysis pipeline only fits rest-frame wavelengths redward of the Lyα line. However, it is well known that the size of the proximity zone and the strength of the damping wing are sensitive to the quasar lifetime t_Q (Bolton & Haehnelt 2007b; Bolton et al. 2011; Keating et al. 2015; Eilers et al. 2017a; Davies et al. 2018c), which can vary from $10^4 - 10^8$ yr (Eilers et al. 2017a, 2021; Khrykin et al. 2021), and thus must be treated as a nuisance parameter. Davies et al. (2018b) presented the first complete model of the proximity zone and damping wing region of quasar spectra by combining the Davies et al. (2018c) estimator for the intrinsic quasar continuum and its associated uncertainty, with a model for the small-scale density fluctuations in the IGM, a description of the reionization topology surrounding the massive dark matter haloes hosting quasars (Davies & Furlanetto 2022), and time-dependent ionizing photon RT (Davies, Furlanetto & McQuinn 2016). To date, the Davies et al. (2018b) modelling approach has been applied to four $7 \lesssim z \lesssim 7.5$ quasars (Bañados et al. 2018; Davies et al. 2018c; Wang et al. 2020; Yang et al. 2020) yielding robust constraints on $\langle x_{\mathrm{HI}}(z) \rangle$ that are competitive with the CMB. Furthermore, the two independent modelling pipelines of Greig et al. (2017b) and Davies et al. (2018c), which adopt distinct approaches for treating the impact of the quasar radiation, reionization topology, and the intrinsic quasar continuum yield results in very good agreement (Greig et al. 2022). Progress on analysing larger samples of quasars, and specifically

$z < 7$ quasars for which $\langle x_{\mathrm{HI}}(z) \rangle$ should be lower, has recently been made by Ďurovčíková et al. (2024) who analysed stacks of a sample of 18 quasars at $6 \lesssim z \lesssim 7$, and by Greig et al. (2024a) who applied the Greig et al. (2017b) pipeline to a sample of 42 quasar spectra at $5.8 \lesssim z \lesssim 6.6$.

These IGM damping wing analyses show that, after marginalizing out nuisance parameters describing the quasar's intrinsic continuum and lifetime, each quasar measures $\langle x_{\mathrm{HI}}(z_{\mathrm{QSO}}) \rangle$ to $\sim 10 - 25$ per cent precision at the quasar redshift, z_{QSO} (see also Kist, Hennawi & Davies 2025). Higher precision constraints on $\langle x_{\mathrm{HI}}(z) \rangle$ thus require averaging independent measurements over large statistical samples of quasars. The recently launched ESA/*Euclid* satellite is poised to discover over 100 quasars with $7.0 \lesssim z \lesssim 7.5$, and ~ 25 quasars beyond the current record of $z = 7.6$, including ~ 8 beyond $z = 8.0$ (Euclid Collaboration 2019). The *JWST* will obtain exquisite spectra of these *Euclid* EoR quasars, many of which will be too faint to be observed with ground-based telescopes in a reasonable observing allocation. Indeed, five existing $z > 6.8$ quasars have already been observed by *JWST* (e.g. Christensen et al. 2023). The combination of large *Euclid* quasar samples and sensitive *JWST* spectra have the potential to revolutionize the study of IGM damping wings towards quasars and constrain the cosmic reionization history to unprecedented precision.

But exploiting the tremendous potential of these facilities to yield precise measurements of $\langle x_{\mathrm{HI}}(z) \rangle$ requires the quantitative study of reionization using quasar damping wings to be on the same solid methodological and statistical footing as other precision cosmological measurements. It is well known that the CMB electron scattering optical depth, τ_e, is degenerate with other cosmological parameters such as the *Hubble* constant, H_0, the amplitude of matter fluctuations, σ_8, or the sum of neutrino masses Σm_ν. Growing tensions between CMB determinations of these parameters and their values measured from late-time probes such as baryon acoustic oscillations (e.g. DESI Collaboration 2025), weak-lensing (e.g. Hikage et al. 2019; van den Busch et al. 2022; Abbott et al. 2022), the cosmic distance ladder (e.g. Riess et al. 2022; Murakami et al. 2023), and laboratory beta-decay experiments (e.g. Di Valentino & Melchiorri 2022; Kreisch, Cyr-Racine & Doré 2020) has spawned an industry of research on these putative anomalies. It is telling that this enormous body of work only uses Planck Collaboration VI (2020) reionization constraints because measurements from the spectra of distant astronomical objects are, apparently, not yet considered credible.

The goal of this paper is to introduce a framework that will elevate the study of IGM damping wings towards quasars to be a precision cosmological probe of reionization. The most significant limitation of past IGM damping wing studies is that they are fundamentally suboptimal for two reasons. First, the commonly adopted red–blue prediction approach forgoes the additional continuum constraining power afforded by spectral pixels blueward of the typically chosen dividing line of 1280 Å. But these blue spectral pixels ($\lambda < 1280$ Å) contain an abundance of information about the intrinsic quasar continuum and the astrophysical parameters $\langle x_{\mathrm{HI}} \rangle$ and t_Q. The most intuitive way to see this is that red–blue continuum prediction provides no mechanism to explicitly prevent the blue-side continuum from lying several spectral noise standard deviations, σ_λ, below the observed spectrum, f_λ, for large swaths of pixels, although such continua are clearly unphysical. Furthermore, jointly fitting the spectral range $\lambda < 1280$ Å, where the smooth damping wing absorption is imprinted, for both the continuum and the astrophysical parameters, $\langle x_{\mathrm{HI}} \rangle$ and t_Q, will surely help break the degeneracy between damping wing strength and intrinsic continuum shape. Second, nearly all past work has not modelled the highly absorbed

proximity zone region blueward of the Lyα line ($\lambda \lesssim 1216\,\text{Å}$), owing to the dual challenge of modelling both the impact of the quasar radiation and the large correlated IGM transmission fluctuations at these wavelengths. Exceptions are Davies et al. (2018b) who performed ionizing RT (see also Bolton et al. 2011) and used simulation-based inference to approximate the intractable likelihood for the correlated transmission fluctuations, and Ďurovčíková et al. (2024), who similarly performed RT and modelled the correlated transmission fluctuations with a Gaussian likelihood – accurate only because they analysed stacked and hence effectively Gaussianized spectra (although stacking is suboptimal and significantly degrades the precision on $\langle x_{\text{HI}}(z) \rangle$[3]). But an accurate likelihood that allows one to fit these absorbed proximity zone pixels would clearly provide information about both the astrophysical parameters and the underlying continuum, since the former determines the distribution of IGM transmission fluctuations, which in turn constrains the latter in a statistical sense (and vice versa). Finally, to be taken seriously as a precision cosmological probe, one must use mock observations to establish that the measurements and the parameter uncertainties quoted are reliable, which has to date never been undertaken.

Building upon the modelling approach presented in Davies et al. (2018b, c), we present an improved technique for constraining reionization with EoR quasar spectra and establish the statistical robustness of the inferred astrophysical parameter constraints. Whereas past work failed to exploit the full constraining power of the blue side spectral region, our key innovation is the construction of a single Bayesian likelihood for the entire spectrum, allowing us to fit the continuum and the IGM damping wing signature simultaneously. We use dimensionality reduction (DR) to describe the intrinsic quasar continuum with a parametric model, but in contrast with most previous IGM damping wing work there is no red–blue prediction, but rather a single latent variable model for the entire quasar spectrum. Using a large training set of 15 559 SDSS/BOSS quasars at $z \gtrsim 2.15$, we trained and quantified the performance of six distinct DR methods, including machine-learning approaches, and find that a six parameter PCA model (five PCA coefficients plus a normalization) performs best, with complex machine learning providing no improvements in performance. All sources of error – the stochasticity induced by the ionization topology, the unknown QSO lifetime t_{Q}, continuum reconstruction errors, and spectral noise – are accounted for in a principled manner, allowing us to marginalize out all continuum nuisance parameters. Finally, by conducting statistical inference on 100 realistic mock EoR quasar spectra, we show that our posterior distributions pass a coverage test, establishing the reliability of the credibility contours that we obtain on $\langle x_{\text{HI}}(z) \rangle$ and t_{Q} from this new method.

In our companion paper (Kist et al. 2025), we quantify the precision with which IGM damping wings analysed with this new inference approach can measure the astrophysical parameters, $\langle x_{\text{HI}} \rangle$ and $\log_{10}(t_{\text{Q}}/\text{yr})$, and the dependence of this precision on the dimensionality of the DR latent variable model, as well as on the spectral resolution, S/N, and spectral coverage of the quasar spectra that are analysed.

The structure of this paper is as follows. In Section 2, we derive the expression for the likelihood of the quasar spectrum that is at the heart of the technique. An exploration of the six DR methods and a description of the training data and procedure are the subject of

Section 3. In Section 4, we summarize the Davies et al. (2018b) approach for simulating quasar transmission spectra with IGM damping wing absorption and quasar lifetime effects, describe how these simulations are used to determine the ingredients required for the quasar spectrum likelihood, and explain our procedure for creating mock quasar spectra. In Section 5, we describe our Hamiltonian Monte Carlo (HMC) based statistical inference, show examples of the inference on mock spectra, present the procedure and results from the coverage testing, and compare the accuracy of our continuum reconstructions to previous work based on red–blue prediction. Finally, we summarize and conclude in Section 6. Appendix A introduces the formalism behind our coverage tests as well as a novel approach to reweight the HMC parameter samples to guarantee that we pass a coverage test even if our original posterior distributions are overconfident.

In this work, we assume a flat Lambda-cold dark matter cosmology with $h = 0.685$, $\Omega_b = 0.047$, $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$, and $\sigma_8 = 0.8$.

2 FORMALISM

Our goal is to derive an expression for the likelihood of a quasar spectrum $\boldsymbol{f}$ with noise vector $\boldsymbol{\sigma}$ (with elements f_λ and σ_λ, respectively) to be observed in a possibly neutral IGM. If we unrealistically assume perfect knowledge of the IGM transmission field, $\boldsymbol{t}$, and the underlying unabsorbed quasar spectrum, $\boldsymbol{s}$, then because the spectral noise is Gaussian distributed, the probability of measuring $\boldsymbol{f}$ is simply

$$P(\boldsymbol{f}|\boldsymbol{\sigma}, \boldsymbol{t}, \boldsymbol{s}) = \mathcal{N}(\boldsymbol{f}; \boldsymbol{t} \circ \boldsymbol{s}, \boldsymbol{\Sigma}), \tag{1}$$

where $\mathcal{N}(\boldsymbol{f}; \boldsymbol{\mu}, \boldsymbol{K})$ is the standard normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{K}$, $\circ$ represents an element wise (Hadamard) product of vectors, and $\boldsymbol{\Sigma} \equiv \text{diag}(\boldsymbol{\sigma})$ is the diagonal matrix formed from the measured noise vector $\boldsymbol{\sigma}$ (throughout we denote vectors with bold lowercase letters/symbols and matrices with bold capital letters/symbols).

The intrinsic quasar spectrum $\boldsymbol{s}$ and the IGM transmission $\boldsymbol{t}$ are latent stochastic processes, which is to say that they are random variables that are not directly observable. Instead, they are related to observables, but we must marginalize over their probability distributions in order to measure the astrophysical parameters of interest. We will adopt this approach in what follows.

First consider the intrinsic quasar spectrum, $\boldsymbol{s}$ – our knowledge is clearly limited by spectral noise, and, blueward of rest-frame Lyα, by IGM absorption. Furthermore, quasar spectra cannot be modelled from first principles. As a result it is common to adopt a data-driven approach and describe $\boldsymbol{s}$ with a DR algorithm, of which PCA (e.g. Suzuki et al. 2005; Suzuki 2006; Pâris et al. 2011; Davies et al. 2018c) is the simplest example. This results in a parametric model, $\boldsymbol{s}_{\text{DR}}(\boldsymbol{\eta})$, where $\boldsymbol{\eta}$ is a new latent variable describing the unabsorbed quasar spectrum which lives in a space with dimensionality lower than the number of spectral pixels (i.e. wavelengths) in $\boldsymbol{s}$. For example, in a PCA decomposition, $\boldsymbol{\eta}$ would be the vector of PCA coefficients plus an overall normalization parameter. Since the unabsorbed spectral pixels redward of the Lyα line provide considerable information about $\boldsymbol{\eta}$, it is advantageous to fit for these parameters, rather than completely marginalize over the quasar continuum stochastic process $\boldsymbol{s}$.

However, no DR algorithm is perfect, which motivates defining the relative reconstruction error

$$\boldsymbol{\delta} \equiv \frac{\boldsymbol{s} - \boldsymbol{s}_{\text{DR}}(\boldsymbol{\eta})}{\boldsymbol{s}}, \tag{2}$$

---

[3]The constraints from the Ďurovčíková et al. (2024) stacked spectra are scarcely more precise than those from individual quasars obtained by other workers (Greig et al. 2017b; Davies et al. 2018c; Greig et al. 2022, 2024a).

where division here is understood to be element wise (i.e. analogous to the Hadamard products of vectors defined in equation 1). It then follows that $s = s_{\mathrm{DR}}(\eta) \circ (1 + \delta)$ to lowest order in δ. We assume that δ is a stochastic process that follows a Gaussian probability distribution function (PDF) given by

$$\mathcal{N}(\delta; \langle \delta \rangle, \Delta), \tag{3}$$

where $\langle \delta \rangle$ and Δ are the mean and covariance matrix of the relative reconstruction error, δ, respectively, which can be empirically determined by applying the DR algorithm to a 'test' data set. We carry out this procedure in Section 3, where it is shown that a Gaussian form is indeed a very good approximation. Given these assumptions, we can finally write for the PDF of the latent variable s:

$$P(s|\eta) = \mathcal{N}(s; \langle s(\eta) \rangle, C_s(\eta)), \tag{4}$$

where we have defined $\langle s(\eta) \rangle \equiv s_{\mathrm{DR}}(\eta) \circ (1 + \langle \delta \rangle)$, $C_s(\eta) = S_{\mathrm{DR}}(\eta) \Delta S_{\mathrm{DR}}(\eta)$, and $S_{\mathrm{DR}}(\eta) \equiv \mathrm{diag}(s_{\mathrm{DR}}(\eta))$.

Next, consider the stochastic process governing the IGM transmission t. While noise and the unknown continuum, s, also limits our knowledge of this latent variable, the primary source of stochasticity is 'cosmic variance', resulting from the unknown initial conditions of the Universe. To make this more concrete, consider how realizations of t are generated. A common approach is to post-process cosmological hydrodynamical simulation outputs with ionizing RT (Bolton & Haehnelt 2007a; Davies et al. 2016, 2018b; Chen & Gnedin 2021; Satyavolu et al. 2023; Zhou et al. 2024), as we discuss further in Section 4. As such, the IGM transmission, $t(\phi, \theta)$ depends on a vector, ϕ, which are the random phases and amplitudes of Gaussian distributed complex latent variables used to initialize the simulation, which represents the unknown initial conditions of the Universe in the vicinity of the quasar, as well as a vector of astrophysical parameters θ, which for IGM damping wing analysis, would be the average IGM neutral fraction $\langle x_{\mathrm{HI}} \rangle$ and the quasar lifetime t_Q, in the simplest description (e.g. Davies et al. 2018b). Our aim is to measure θ, whereas it is clearly computationally intractable to attempt to fit for the latent variables ϕ, unless one had an incredibly fast way of simulating the IGM and performing the RT. As it is unclear whether fitting for ϕ would be advantageous (but see Horowitz et al. 2019) and it is clearly computationally intractable, the obvious strategy is to marginalize over the initial conditions, ϕ.

Based on the foregoing considerations, we perform the marginalization over the relative reconstruction error, δ, and the initial conditions, ϕ, to arrive at an expression for the likelihood of the data f given observed noise vector σ, astrophysical model parameters θ, and DR latent variables η:

$$L(f|\sigma, \theta, \eta) = \iint P(f|\sigma, t(\phi, \theta), s) P(s|\eta) P(\phi) \mathrm{d}s \mathrm{d}\phi. \tag{5}$$

Using the definition of the Dirac delta function

$$1 = \int \delta_{\mathrm{D}}[t - t(\phi, \theta)] \, \mathrm{d}t, \tag{6}$$

we can introduce an additional integral over the stochastic variable t in equation (5) giving

$$L(f|\sigma, \theta, \eta)$$
$$= \iiint P(f|\sigma, t, s) P(s|\eta) \delta_{\mathrm{D}}[t - t(\phi, \theta)] P(\phi) \mathrm{d}s \mathrm{d}\phi \mathrm{d}t. \tag{7}$$

The probability distribution of IGM transmission, $P(t|\theta)$, can now be defined as

$$P(t|\theta) = \int \delta_{\mathrm{D}}[t - t(\phi, \theta)] P(\phi) \mathrm{d}\phi, \tag{8}$$

which is the pushforward distribution of the prior over initial conditions, $P(\phi)$, through the deterministic mapping $t(\phi, \theta)$. Intuitively, $P(t|\theta)$ describes the distribution of IGM transmission fields that results from drawing random initial conditions ϕ according to $P(\phi)$ and evolving them forward under the simulation specified by the astrophysical parameters θ. Finally, we can write the likelihood as

$$L(f|\sigma, \theta, \eta) = \iint P(f|\sigma, t, s) P(s|\eta) P(t|\theta) \mathrm{d}t \mathrm{d}s. \tag{9}$$

Note that two of the PDFs under the integral in equation (9), $P(f|\sigma, t, s)$ and $P(s|\eta)$ have a Gaussian form, but $P(t|\theta)$ poses a challenge since, whether at random locations in the Universe (e.g. Lee et al. 2015; Davies et al. 2018a) or in quasar proximity zones (Davies et al. 2018b), it is well known that the PDF of the IGM transmission is non-Gaussian. While generating samples from $P(t|\theta)$ is straightforward – simply randomly select IGM transmission skewers generated from a simulation with parameters θ – there exists no tractable analytical expression for $P(t|\theta)$. It thus follows that it is impossible to derive an exact analytical expression for the desired likelihood both because $P(t|\theta)$ is intractable, and because even if an expression for it existed, it would be extremely challenging to perform the high-dimensional marginalization integrals in equation (9).

Our approach going forward is to approximate $P(t|\theta)$ with a Gaussian form,

$$P(t|\theta) = \mathcal{N}(t; \langle t(\theta) \rangle, C_t(\theta)), \tag{10}$$

where $\langle t(\theta) \rangle$ and $C_t(\theta)$ are the mean IGM transmission and its covariance, which are easily measured from realizations using forward simulations of quasar proximity zones. With this approximation, we can now obtain an approximate analytic expression for the likelihood. Substituting the Gaussian PDFs for f, s, and t from equations (1), (4), and (10) into the marginalization integral in equation (9) gives

$$L(f|\sigma, \theta, \eta) = \iint \mathcal{N}(f; t \circ s, \Sigma) \mathcal{N}(s; \langle s(\eta) \rangle, C_s(\eta))$$
$$\times \mathcal{N}(t; \langle t(\theta) \rangle, C_t(\theta)) \mathrm{d}s \mathrm{d}t. \tag{11}$$

Re-arranging to express the integral over t in terms of the variable $t \circ s$ gives

$$L(f|\sigma, \theta, \eta) = \int \left[\int \mathcal{N}(f; t \circ s, \Sigma) \mathcal{N}(t \circ s; \langle t \rangle \circ s, S C_t S) \mathrm{d}(t \circ s) \right]$$
$$\times \mathcal{N}(s; \langle s \rangle, C_s) \mathrm{d}s, \tag{12}$$

where $S \equiv \mathrm{diag}(s)$ and we suppress explicit dependencies of the means and covariances on θ and η for notational brevity. Exploiting the fact that Gaussians are closed under convolution, the integral in brackets can be analytically evaluated giving

$$L(f|\sigma, \theta, \eta) = \int \mathcal{N}(f; \langle t \rangle \circ s, \Sigma + S C_t S) \mathcal{N}(s; \langle s \rangle, C_s) \mathrm{d}s. \tag{13}$$

Analogous to above, we can re-arrange equation (13) in terms of the variable $\langle t \rangle \circ s$ giving

$$L(f|\sigma, \theta, \eta) = \int \mathcal{N}(f; \langle t \rangle \circ s, \Sigma + S C_t S)$$
$$\times \mathcal{N}(\langle t \rangle \circ s; \langle t \rangle \circ \langle s \rangle, \langle T \rangle C_s \langle T \rangle) \mathrm{d}(\langle t \rangle \circ s). \tag{14}$$

where $\langle T \rangle \equiv \mathrm{diag}(\langle t \rangle)$. While the form of equation (14) suggests one use the closure of Gaussians under convolutions again to evaluate the integral over $\langle t \rangle \circ s$, note that s also now appears in the covariance of the first normal distribution via the term $S C_t S$ (recall that $S \equiv \mathrm{diag}(s)$), which instead renders this integral intractable. To make progress we approximate

$$s = s_{\mathrm{DR}}(\eta) \circ (1 + \delta) \approx s_{\mathrm{DR}}(\eta) \circ (1 + \langle \delta \rangle) \equiv \langle s(\eta) \rangle \tag{15}$$

in the problematic covariance term such that

$$SC_tS \approx \langle S \rangle C_t \langle S \rangle \qquad (16)$$

where $\langle S \rangle \equiv \mathrm{diag}(\langle s(\eta) \rangle) = \mathrm{diag}[s_{\mathrm{DR}}(\eta) \circ (1 + \langle \delta \rangle)]$. This removes the s dependence from the covariance[4] so that the likelihood becomes

$$
\begin{aligned}
L(f|\sigma, \theta, \eta) &= \int \mathcal{N}(f; \langle t \rangle \circ s, \Sigma + \langle S \rangle C_t \langle S \rangle) \\
&\quad \times \mathcal{N}(\langle t \rangle \circ s; \langle t \rangle \circ \langle s \rangle, \langle T \rangle C_s \langle T \rangle) \mathrm{d}(\langle t \rangle \circ s). \\
&= \mathcal{N}(f; \langle t \rangle \circ \langle s \rangle, \Sigma + \langle S \rangle C_t \langle S \rangle + \langle T \rangle C_s \langle T \rangle), \quad (17)
\end{aligned}
$$

where the last equality again follows from the closure of Gaussians under convolution.

The primary virtue of the likelihood in equation (17) is that it operates on the entire quasar spectrum, and constitutes a significant departure from the now standard approach in quasar IGM damping wing analysis of using the spectrum redward of the Lyα line to predict the intrinsic spectrum bluward of Lyα, and then performing inference on the resulting normalized spectrum (e.g. Davies et al. 2018b; Ďurovčíková et al. 2020; Fathivavsari 2020; Reiman 2020; Chen et al. 2022). Clearly pixels bluward of the typical red–blue dividing line of $\simeq 1280$ Å contain an abundance of information about the intrinsic spectrum and the astrophysical parameters θ. As it operates on the entire spectrum, the likelihood in equation (17) fully incorporates information from pixels with $\lambda < 1280$ Å, where the smooth damping wing absorption is imprinted. Fitting this region helps break the degeneracy between damping wing strength and intrinsic spectrum shape. Finally, even partly absorbed pixels in the proximity zone constrain $s_{\mathrm{DR}}(\eta)$, since each model θ predicts the distribution of t, which in turn constrains $s_{\mathrm{DR}}(\eta)$ in a statistical sense. Because $t_\lambda = \exp(-\tau_\lambda)$ and $s_\lambda = f_\lambda/t_\lambda$, the uncertainty in the continuum arising from stochastic IGM fluctuations $|\delta s_\lambda / s_\lambda| = |\delta t_\lambda/t_\lambda| = |\delta \tau_\lambda|$ will be the smallest at low optical depth, and hence the highest transmission, t, inner proximity zone pixels (i.e. closest to the Lyα wavelength 1215.67 Å in the rest frame) arising from regions illuminated by the quasars intense radiation will contain the most information about the intrinsic spectrum s. This also suggests that observations with resolution sufficient to spectrally resolve transmission spikes in the proximity zone could afford additional constraining power (but see Kist et al. 2025).

The main disadvantage of the likelihood in equation (17) is that it is approximate, with the most significant errors incurred from assuming a Gaussian form for $P(t|\theta)$ in equation (10), which we investigate in detail in Section 5.6. In Section 5, we show that this approximation yields biased and overconfident parameter inference, but a strategy for mitigating these shortcomings is introduced. In our companion paper, we conduct more detailed tests of the precision and fidelity of the statistical inference delivered by equation (17) and better understand the conditions under which the Gaussian approximation for the PDF of the proximity zone transmission is valid (Kist et al. 2025).

Finally, we note that our approach bares some resemblance to the likelihoods derived by Garnett et al. (2017) and Sun, Ting & Cai (2023) in the context of the lower-z Lyα forest. However, our

---

[4] This approximation suppresses the modulation of the covariance by the fluctuations due to the relative reconstruction error δ. While these fluctuations are small, ignoring them is not obviously mathematically justifiable in terms of an expansion in powers of δ. But given the already crude approximation of a Gaussian transmission PDF (i.e. equation 10), this inconsistency is tolerable as it yields a closed form analytical expression.

likelihood is more accurate since both Garnett et al. (2017) and Sun et al. (2023) incorporate IGM absorption via an additive noise term. Although it simplifies the math, this approximation surely breaks down for the highly opaque IGM of interest to us here, whereas our analysis treats IGM absorption as multiplicative and is both more accurate and applicable to low-z and high-z IGM absorption alike. Furthermore, Garnett et al. (2017) assume that the intrinsic quasar spectrum, s in our notation, follows a Gaussian distribution, whereas Sun et al. (2023) assume that the latent variables describing the intrinsic spectrum, η in our notation, are Gaussian distributed and that the covariance of the continuum reconstruction errors, $C_s(\eta)$ in our notation, is a diagonal matrix. In contrast, we only assume Gaussianity for the relative reconstruction error, δ, which is a far weaker assumption, and treat fully covariant continuum reconstruction errors (see Fig. 7). The main advantage of their approaches relative to ours is that they present a method to determine the quasar continuum DR model, $s_{\mathrm{DR}}(\eta)$ directly from a low-z Lyα forest data set in addition to the astrophysical parameters, whereas in our approach, we derive the latent variable model from an external training set, which as we discuss in the next section, comes from SDSS/BOSS $2.15 \lesssim z \lesssim 4$ quasar spectra fit with an automated continuum fitting algorithm. It is worth exploring in future work if, analogous to Garnett et al. (2017) and Sun et al. (2023), one can use the likelihood in equation (17) to fit for both the latent variable model that describes the continua of SDSS/BOSS $2.15 \lesssim z \lesssim 4$ quasar spectra and the astrophysical parameters that govern the low-z Lyα forest.

3 QUASAR DIMENSIONALITY REDUCTION

A critical component of the formalism presented in the previous section is the representation of the intrinsic quasar spectrum, s, with a DR algorithm. Specifically, before we can compute the likelihood in equation (17) we need to: (1) apply a DR algorithm to an ensemble of quasar spectra to determine the function $s_{\mathrm{DR}}(\eta)$, (2) demonstrate that the probability distribution of the relative reconstruction error δ (equation 2) is well described by a multivariate Gaussian distribution (equations 3 and 4), and (3) measure the mean $\langle \delta \rangle$ and covariance Δ of this distribution.

An important question is which DR algorithm to employ. We will compare several different DR approaches in this section. DR methods are commonly divided into linear and non-linear models. We will start with most widely adopted linear model which is PCA, which will be compared to two non-linear models, namely a Gaussian Process Latent Variable Model (GPLVM), and a variational autoencoder (VAE). Although PCA is provably optimal among linear methods in the sense that it minimizes the average-squared reconstruction error for a fixed dimension of the latent space, n_{latent}, this optimality does not extend to non-linear methods. In particular, non-linear DR algorithms such as the GPLVM and VAE could, in principle, more efficiently capture the structure of the data manifold, and thus achieve smaller reconstruction errors. For this reason, we assess the performance of each method empirically. First, we will describe the procedure for generating the training data for the DR algorithms (Section 3.1), then we will discuss our implementation of each DR method (Section 3.2), and finally we will quantify and compare their performance for representing quasar spectra (Section 3.3). Ultimately, we will conclude that a PCA with $n_{\mathrm{DR}} = 6$ parameters (i.e. $n_{\mathrm{latent}} = 5$ PCA coefficients plus a normalization parameter, s_{norm}) is the best choice for our application. The reader who is not interested in the details can skip ahead to Section 3.4.

3.1 Training and test data

3.1.1 Automated continuum fits of SDSS quasars

Our training set of quasar spectra is drawn from the SDSS-III BOSS and the SDSS-IV Extended BOSS (eBOSS) surveys which obtained moderate resolution ($R \sim 2000$) spectra of a large sample of $z \gtrsim 2$ quasars (Dawson et al. 2013, 2016). Specifically, we consider objects identified as quasars in the eBOSS DR14 quasar catalogue (Pâris et al. 2018) using the compilation in the igmspec database of public spectra probing the IGM, which uses the specdb database framework (Prochaska 2017). Our aim is to use DR to describe quasars over the rest-frame wavelength range relevant to IGM damping wing analysis 1170–2040 Å. As the (upgraded) SDSS spectrograph covers the wavelength range 3580–10 350 Å, the requirement that our desired rest-frame wavelength range be fully covered restricts the range of usable quasar redshifts. Adding a small buffer on the blue end of the spectra to avoid edge effects, and excluding quasars with $z > 4$ where the Lyα forest transmission is smaller and harder to correct with automated continuum fitting algorithms (see below), we arrive at the redshift range $2.149 < z < 4.0$. There are 199 530 quasars in this range in the eBOSS DR14 quasar catalogue, of which we removed 13 391 objects that are likely broad absorption line (BAL) quasars given non-zero values of the BALnicity index characterizing C IV absorption troughs (see e.g. Pâris et al. 2018), resulting in a sample of 186 139 quasars. We further require that the median signal-to-noise ratio, S/N > 10 within a 5.0 Å region centred at rest-frame 1285 Å. After imposing this S/N requirement and removing a small number of problematic spectra with large gaps in their spectral coverage, we are left with a parent sample of 20 201 quasars.

In order to define a continuous smooth spectrum that covers our desired rest-frame wavelength range, we use automated continuum fitting algorithms following the approaches adopted by previous work on quasar continua (Davies et al. 2018c; Ďurovčíková et al. 2020; Bosman et al. 2021). Davies et al. (2018c) and Bosman et al. (2020) used the automated fitting procedure developed by Young et al. (1979) and Carswell et al. (1982) as implemented by Dall'Aglio, Wisotzki & Worseck (2008), which determines a smooth continuum in the presence of absorption lines and noise both blueward and redward of the Lyα emission line. This algorithm iteratively fits the spectra with a cubic spline with breakpoints initially spaced by ~ 1400 km s^{-1} in the Lyα forest (i.e. 20 SDSS pixels, each pixel is 70 km s^{-1}) and ~ 1100 km s^{-1} redward of the Lyα line (16 pixels). Pixels that lie more than two standard deviations below the fit are iteratively rejected. Additional spline breakpoints are added if the slope between neighbouring breakpoints exceeds a threshold, and spline points are merged if the variations between neighbouring breakpoints are small. We applied this algorithm to all 20 201 quasars in our target redshift range, which we will henceforth refer to as the autofit continua. For simplicity, our DR does not attempt to capture luminosity-dependent changes in quasar spectral shape. As such, DR will work best if all quasars are on a common flux scale, and we thus rescale each of the autofit continua to be unity at a rest wavelength of 1285 Å and rescale each of the training data spectra by the same factor.

Ďurovčíková et al. (2020) followed a different approach to automatically fit continua to SDSS spectra, which we will refer to as the QSMOOTH continua. Briefly, they first compute a running median with a width of 50 SDSS spectral pixels to capture the salient continuum and emission features in the spectrum. Peak finding is then performed on the spectrum with the requirement that the peaks lie above a local threshold set by this running median spectrum, and these peaks are spline interpolated to define an upper envelope for

the spectrum. After subtracting this envelope from the spectrum, the RANSAC regression algorithm is applied to the residuals to define inliers (the continuum level) and outliers (the absorption lines). The data points that are identified as inliers are interpolated and smoothed by computing a running median with a bin size of 20 pixels, resulting in the final automated continuum fit to the spectrum. We apply a slightly modified version of the publicly available QSMOOTH[5] code to the 20 201 quasars in our target redshift range. Note that the QSMOOTH algorithm was applied to the quasar spectra after rescaling them such that their autofit continua equal unity at 1285 Å, so we do not independently renormalize the QSMOOTH continua.

Once we have these two independent estimates of the quasar continuum, we use them to attempt to further remove problematic objects from the training data. Associated absorption around the Lyα and N V emission-line complex resulting from either BAL absorption, proximate DLA or Lyman limit systems (PDLAs or PLLSs), or strong metal absorption due to proximate N V or intervening absorbers, will result in artefacts in the automated continuum fits. We remove these cases by simply discarding objects for which either the autofit or QSMOOTH continua fall below 0.6 in the wavelength range 1170 Å $< \lambda < 1285$ Å (recall that the spectra are normalized to unity at 1285 Å). In a similar spirit, we attempt to also exclude BALs by removing any objects which have either of their automated continua < 0.1 in the wavelength range 1285 Å $< \lambda < 1990$ Å, or which have automated continua < 0.7 in the wavelength range 1300 Å $< \lambda < 1570$ Å. These thresholds and wavelength ranges were all determined via trial and error, and collectively these cuts remove 4433 objects, leaving 15 768 quasars.

The SDSS spectra are reported on their native observed frame wavelength grid, but the DR analysis requires a common rest-frame wavelength grid. We convert the observed wavelengths to rest wavelengths by dividing by $1 + z_{QSO}$, where the quasar redshift, z_{QSO} is taken to be the pipeline redshift, Z_PIPE, in the DR14 quasar catalogue. We construct a rest-frame wavelength grid that is linearly spaced in velocity (or equivalently $\log_{10} \lambda$) with a pixel size of $dv_{pix} = 140$ km s^{-1} (roughly twice the native SDSS pixel scale) spanning from $1170 - 2040$ Å, resulting in a total of $n_{spec} = 1190$ spectral pixels. The spectra are rebinned onto this wavelength grid by averaging the flux values of the spectral pixels that land in a given wavelength bin. The autofit and QSMOOTH continua are interpolated onto this wavelength grid.

A small fraction of objects in the eBOSS DR14 quasar catalogue have incorrect redshifts. We found that an effective way to automatically identify and remove these objects is to analyse their Lyα forest transmission. Specifically, we use the autofit continua to normalize the rebinned SDSS spectra and compute the mean Lyα forest transmission of each object in the wavelength range 1170 Å $< \lambda < 1190$ Å. We compare this measured mean flux $\langle F \rangle$ to the empirically determined value $\langle F \rangle_{true}$, where the latter is obtained by evaluating the fit from Oñorbe, Hennawi & Lukić (2017) at the average Lyα absorption redshifts, $z_{Ly\alpha}$, probed by the quasar spectrum. For the redshift range, we consider the Oñorbe et al. (2017) fit is anchored by the Becker et al. (2013) mean flux measurements. For each spectrum a flux contrast can then be defined as $\delta_F = (\langle F \rangle - \langle F \rangle_{true})/\langle F \rangle_{true}$. We then use the median absolute deviation[6] (MAD) to estimate the effective standard deviation, σ_{δ_F}, of δ_F, from all of the quasars which allows us to define $\chi_{\delta_F} \equiv \delta_F/\sigma_{\delta_F}$.

---

[5]https://github.com/DominikaDu/QSmooth
[6]Because occasional large outliers, in this section and in other places in the paper we estimate standard deviations from the MAD using $\sigma = 1.4826 \times$ MAD.

Empirically, we find that large positive fluctuations χ_{δ_F} correspond to spurious redshifts, which are typically lower redshift quasars which do not exhibit Lyα forest absorption in their SDSS spectra because another emission line has been incorrectly identified as Lyα. From trial and error we find that applying a cut of $\chi_{\delta_F} > 5$ removes the vast majority of misidentified redshifts, which removes 54 objects from our sample.

Finally, we found that the difference between the autofit and the QSMOOTH continua provides a reasonable proxy for the continuum fitting 'error'. If we define $\Delta s_{i\lambda} \equiv s_{\text{autofit},i\lambda} - s_{\text{QSmooth},i\lambda}$, where i is the index of the quasar in question and λ is the wavelength, then the inverse variance of the continuum fit can be defined is

$$\frac{1}{\sigma_{s,i\lambda}^2} \equiv \frac{1}{\Delta s_{i\lambda}^2} + \frac{\left|s_{\text{autofit},i\lambda}\right|^2}{\text{S/N}_{\text{clip}}^2}, \tag{18}$$

where S/N$_{\text{clip}}$ applies a flux-dependent floor to the noise that guarantees that the signal-to-noise ratio, S/N $\equiv s_{\text{autofit},i\lambda}/\sigma_{s,i\lambda}$ never exceeds S/N$_{\text{clip}}$, which we set to be S/N$_{\text{clip}} = 30$. We also define the relative fluctuation $\delta s_{i\lambda} \equiv \Delta s_{i\lambda}/s_{\text{autofit},i\lambda}$, and analogous to above, quantify fluctuations relative to the typical scatter via

$$\chi_{s,i\lambda} \equiv \frac{\delta s_{i\lambda} - \langle\delta s_\lambda\rangle}{\sigma_{\delta s,\lambda}}, \tag{19}$$

where the $\langle\delta s_\lambda\rangle$ and $\sigma_{\delta s,\lambda}$ are the wavelength-dependent mean and standard deviation of $\delta s_{i\lambda}$, i.e. these moments are computed by averaging out the quasar dimension i. The rms fluctuation of $\chi_{s,i\lambda}$, across 1170 Å $< \lambda <$ 1240 Å, provides a summary statistic quantifying the relative differences between the two autofit continua for each quasar in the range relevant to IGM damping wings. We require that this rms fluctuation is less than four, which removes another 155 quasars, or about 1 per cent, yielding the final size of our DR quasar sample of 15 559.

We opt to use the autofit continua since they visually appear to be closer to the truth than the QSMOOTH continua. The data thus used to train the DR algorithms are the autofit continua, $s_{\text{autofit},i\lambda}$, and the error $\sigma_{s,i\lambda}$. Fig. 1 shows seven objects randomly chosen from the DR quasar sample. Following standard practice in machine learning, we split this DR sample of 15 559 quasars into a training set of 14 781 objects and a test set of 778 objects, corresponding to a roughly 95 per cent − 5 per cent split. The training set will be used to train our DR algorithms and the test set, which we refer to as our autofit test set, will constitute the unseen data that will be used to quantify performance.

3.1.2 Hand-fit continua

The previous section discussed the automated continuum fitting, but it is also possible (although tedious) to fit quasar continua by hand. Whether this approach results in continua that are more accurate than the aforementioned automated algorithms is unclear. But hand-fit continua provide an additional 'test' data set that can be used to quantify the efficacy of DR algorithms which motivates us to compile a set for this purpose. Pâris et al. (2011) hand fit a set of 78 high signal-to-noise ratio, S/N > 14, quasar spectra in the redshift range $2.82 < z < 3.0$, where the S/N was evaluated at rest-frame 1280 Å. The continua were fit by selecting cubic spline breakpoints at regular intervals and adjusting the height and spacing of these breakpoints. In the Lyα forest, where this procedure is particularly challenging and subjective, the breakpoint heights were chosen to follow the peaks of the flux. We renormalize the 78 Pâris et al. (2011) continua to be unity at 1285 Å (for consistency with our autofit continua).

We augment the Pâris et al. (2011) continua, with continua from the XQ-100 survey (López et al. 2016), which acquired a set of 100 high signal-to-noise-ratio (S/N $\sim$ 33 at rest-frame 1700 Å) echelle spectra of quasars at $3.5 \lesssim z \lesssim 4.5$ using the X-shooter instrument on the Very Large Telescope (VLT). The public data release of XQ-100 provides hand-fit continua for every quasar following an algorithm that is similar in spirit to that used by Pâris et al. (2011). As the public release of the XQ-100 data set provided continuum fits only for the spectra obtained by the individual arms of the X-shooter spectrograph, we restrict attention to the VIS arm, which covers the observed frame 5500 Å $< \lambda <$ 10200 Å. As our mock high-z quasar spectra with IGM damping wings (see Section 4) cover the rest-frame wavelength range 1185 Å $< \lambda <$ 2000 Å, the requirement that this entire range is covered by the VIS arm restricts the number of quasars to be 43, and the redshift range to be $3.65 \leq z \leq 4.09$. We similarly renormalize the XQ-100 continua to be unity at 1285 Å. For the quasar redshifts, z_{QSO}, we used the values adopted by Pâris et al. (2011) and the XQ100 survey. Combining the Pâris et al. (2011) and XQ-100 spectra, we finally arrive at a sample of 121 hand-fit continua, which we henceforth refer to as our hand-fit test data set.

3.2 Dimensionality reduction algorithms

A DR algorithm transforms data from a high-dimensional space into a lower dimensional latent space, such that the lower dimensional representation retains as much information as possible from the higher dimensional process. In more concrete terms, our training set comprises $n_{\text{train}} = 14\,781$ quasars with $n_\lambda = 1190$ spectral pixels per quasar. The smooth appearance of a quasar continuum (see Fig. 1) implies these n_λ spectral pixels are highly correlated, and hence the bulk of the information content can be encapsulted by a vector $\boldsymbol{\eta}$ in a lower dimensional latent space of dimension, n_{DR}, which parametrizes a DR model $s_{\text{DR}}(\boldsymbol{\eta})$. Recall from the discussion in Section 3.1 that we normalized all of our training and test set spectra to unity at 1285 Å which removes the amplitude degree of freedom from the stochastic process s. But since our goal is to eventually fit $s_{\text{DR}}(\boldsymbol{\eta})$ to quasar spectra with arbitrary flux, we add an additional multiplicative normalization parameter, s_{norm}. This implies our DR quasar continuum model will have $n_{\text{DR}} = n_{\text{latent}} + 1$ free parameters, i.e. the normalization and the n_{latent} latent variables that parametrize a latent variable model for the normalized spectra. We define $\boldsymbol{\xi}$ to be a vector whose elements are the parameters in the n_{latent} dimensional latent space that describes the normalized spectra, whereas $\boldsymbol{\eta} \equiv (s_{\text{norm}}, \xi_1, \xi_2, ..., \xi_{n_{\text{latent}}})$, is the vector in the $n_{\text{DR}} = n_{\text{latent}} + 1$-dimensional latent space that describes the non-normalized spectra.

DR methods are commonly divided into linear and non-linear algorithms, and we consider both in this study. We start with the most widely adopted linear model, which is PCA.

3.2.1 Principal component analysis

PCA is a commonly used tool to understand correlations in quasar spectra (e.g. Boroson & Green 1992; Francis et al. 1992; Yip et al. 2004; Suzuki 2006). Suzuki et al. (2005) first proposed PCA as a method to predict the continuum absorbed regions of quasar spectra, which was later improved upon by Pâris et al. (2011). This approach, which uses pixels redward of Lyα to predict the continuum in the Lyα forest, has been used in many IGM absorption studies (e.g. Kirkman et al. 2005; Lee et al. 2012, 2013, 2015; Eilers et al. 2017a, b; Eilers, Hennawi & Davies 2018; Ďurovčíková et al. 2020, 2024;
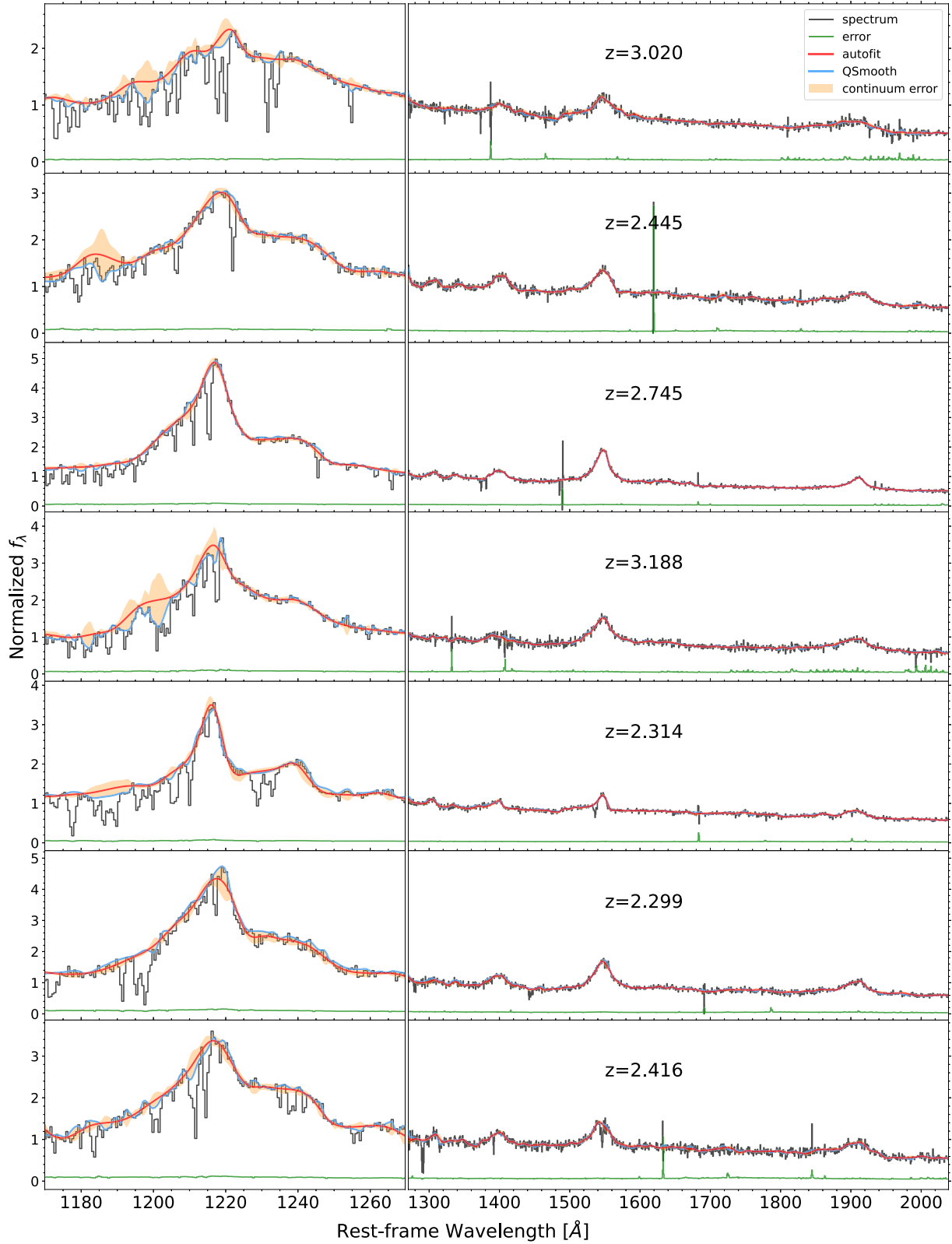
Figure 1. Seven randomly selected quasars from our DR quasar sample. Black and green histograms show the quasar spectrum and 1σ spectral noise, respectively. Automated continuum fits from the autofit and QSMOOTH algorithms are shown in blue and red, respectively. The orange-shaded region indicates the error on the continuum fit derived from the differences between the autofit and QSMOOTH fits, as defined in equation (18). The autofit continua (red) appear to provide good estimates for the continuum and the continuum errors (orange shaded) are generally larger in regions where the continuum is less certain owing to strong absorption.

Bosman et al. 2021). Motivated by this technique, Davies et al. (2018c) developed a PCA-based model from a far larger sample of spectra encompassing a wide range of spectral properties with an eye toward predicting the quasar continua for IGM damping wing analysis. A similar stack was followed by Bosman et al. (2021) who modelled a broader spectral region relevant to the Lyα and Lyβ forests. While the training data and PCA models used here are qualitatively very similar to those used by Davies et al. (2018c) and Bosman et al. (2021), the primary difference is that these works applied PCA to the blue and red spectral regions independently, using the latter to predict the former, whereas here we apply PCA and other DR algorithms to the entire spectrum, since our inference approach models the IGM and the continuum simultaneously using all spectral pixels (see Section 2).

The principal components of a collection of n_{train} quasars that reside in a n_λ dimensional spectrum space can be thought of as a set of unit vectors (with n_λ components) where the ith vector defines the direction of a line in the n_λ dimensional space that best fits the variations among the spectra while being orthogonal to the first $i - 1$ unit vectors. The components are rank ordered according to the amount of variation in the data set that they describe, and typically one truncates the latent space at some value $n_{\text{latent}} < n_\lambda$, such that the dimensionality of the process is reduced.

A quasar PCA decomposition can be written as

$$s_{\text{DR}}(\boldsymbol{\eta}) = s_{\text{norm}} \left(\langle s \rangle + \boldsymbol{\xi}^T A \right), \qquad (20)$$

where $\boldsymbol{\xi}$ is the n_{latent} dimensional latent variable (commonly referred to as the PCA coefficients) and A is the set of n_{latent} PCA vectors or principal components, each of which has n_λ spectral pixels, i.e. $A_{i\lambda}$ is the n_λ dimensional PCA vector corresponding to the ith component.

Davies et al. (2018c) argued that the dominant mode of variations between quasar spectra are their power-law continua, which are more naturally described by PCA decomposition in log space, motivating

$$\ln s_{\text{DR}}(\boldsymbol{\eta}) = \ln s_{\text{norm}} + \langle \ln s \rangle + \boldsymbol{\xi}^T A, \qquad (21)$$

or equivalently

$$s_{\text{DR}}(\boldsymbol{\eta}) = s_{\text{norm}} \left[\exp\left(\langle \ln s \rangle\right) \exp\left(\boldsymbol{\xi}^T A\right) \right], \qquad (22)$$

where $\boldsymbol{\eta} \equiv (s_{\text{norm}}, \xi_1, \xi_2, ..., \xi_{n_{\text{latent}}})$. We will refer to these two decomposition choices as PCA and lnPCA, respectively.

Given that we have an estimate of the errors for our continuum fits (equation 18), we also explore whether there is an advantage to using a weighted PCA.[7] Whereas in standard PCA one performs a singular value decomposition of the sample covariance, weighted PCA instead decomposes a weighted covariance matrix to compute the set of principal component vectors A (Delchambre 2015). For our weighted PCA, we set the weights for each spectral pixel to be the inverse variance defined in equation (18).

To summarize, there are four different variations of PCA that we explore: standard (PCA), log PCA (lnPCA), weighted PCA (wPCA), and weighted log PCA (wlnPCA).

3.2.2 Gaussian process latent variable model

Gaussian processes (GPs) are a supervised learning method for solving regression and classification problems in a powerful non-parametric probabilistic framework. A GPLVM (Lawrence 2005) uses GPs for unsupervised learning tasks like DR or searching for

hidden structure in high-dimensional data. In the current context, the GPLVM will act as a decoder, providing a probabilistic mapping from the latent space variables $\boldsymbol{\xi}$ to the data space variables s. The smoothness of this mapping is controlled by kernel functions whose hyperparameters are fit during the training process. This is analogous to GP regression, where given inputs X and outputs y, one chooses a kernel and learns hyperparameters that best describe the mapping from X to y. In the GPLVM, one is not given the latent variables $\boldsymbol{\xi}$ (i.e. X), but is rather only given s (i.e. y). The latent variables $\boldsymbol{\xi}$ representing each example in the training set must be learned along with the kernel hyperparameters. It is in this sense that the GPLVM acts a decoder, or equivalently a forward mapping from latent space to data space. Once the latent variables that encode each training instance and the kernel hyperparameters are learned, one can evaluate the model, $s_{\text{GPLVM}}(\boldsymbol{\xi})$, at any location in the latent space. This forward mapping is governed by GPs which are independently defined for each dimension of the data space, which is to say that each spectral pixel s_λ has its own GP (and associated hyperparameters) which regresses $\boldsymbol{\xi}$ to produce s_λ. See Eilers et al. (2022) for a recent application of GPLVMs for DR of quasar spectra.

In the canonical formulation of GPLVM (Lawrence 2005), the covariance hyperparameters and point estimates for the unknown latent variables are determined jointly, by optimizing the likelihood of the data. But it is well known that the optimization required for GP regression scales as $\mathcal{O}(n^3)$ for a data set of size n, such that applying canonical GPLVM to 'big data' is computationally prohibitive. Hensman, Fusi & Lawrence (2013) showed how stochastic variational inference (SVI; Hoffman et al. 2013) techniques can be used to apply GPs to very large data sets by stochastically optimizing over mini-batches of the data set. Building upon this, Lalchand, Ravuri & Lawrence (2022) re-formulated the Bayesian incarnation of the GPLVM (Titsias & Lawrence 2010) in an SVI framework by using a structured doubly stochastic lower bound (Titsias & Lázaro-Gredilla 2014) which enables training on very large data sets.

We apply the Lalchand et al. (2022) implementation[8] of Bayesian GPLVM, built in the `gpytorch`[9] GP framework (Gardner et al. 2018), to our quasar training set. The loss function optimized in this formalism exploits the errors on the quasar continua that were defined in equation (18). The training set is passed through a scaler transformation to 'whiten' the data. Namely, the wavelength-dependent mean, $\langle s_\lambda \rangle$, and standard deviation, $\sigma_{s,\lambda}$, of the training set are computed by averaging over the sample of quasars. We rescale the spectra s_λ using the transformation $y_\lambda = (s_\lambda - \langle s_\lambda \rangle)/\sigma_{\text{median}}$, and correspondingly rescale their errors $\sigma_{s,\lambda}$. Here, σ_{median} is a single number which is the median value of $\sigma_{s,\lambda}$.[10]

3.2.3 Variational autoencoder

An autoencoder is an unsupervised learning method that uses an artificial neural network to learn a latent space representation, or

---

[7]We use the WPCA implementation available here: https://github.com/jakevdp/wpca.

[8]https://docs.gpytorch.ai/en/latest/examples/045_GPLVM/Gaussian_Process_Latent_Variable_Models_with_Stochastic_Variational_Inference.html

[9]https://gpytorch.ai

[10]A more common whitening procedure would be to adopt $y_\lambda = (s_\lambda - \langle s_\lambda \rangle)/\sigma_{s,\lambda}$, where $\sigma_{s,\lambda}$ is the standard deviation of each feature (wavelength). However, this transformation changes the shape of the spectra that are fit by the GPLVM by effectively downweighting larger emission-line fluctuations and upweighting smaller continuum variations. We found that scaling by a single wavelength independent number, σ_{median}, yielded better results.

encoding, of a data set. The essential property of autoencoders are their architecture, which consists of an encoder that maps the input into the code, and a decoder that maps the code to a reconstruction of the input. The code or equivalently the latent variables, $\boldsymbol{\xi}$, correspond to the output of a specific layer of a multilayer perceptron (MLP). Typically, the encoder passes the inputs through a sequence of gradually smaller hidden layers of an MLP until hitting a bottleneck which has n_{latent} neurons. The decoder then passes the code back up through gradually larger hidden layers to generate the reconstruction of the input. When the latent space has dimensionality smaller than the data space, autoencoders can be used for DR. A VAE (Kingma & Welling 2013) is similar in spirit to an autoencoder, but it essentially replaces the deterministic decoder, which generates the data from latent variables $\boldsymbol{s}(\boldsymbol{\xi})$, with a probablistic decoder $p(\boldsymbol{s}|\boldsymbol{\xi})$, and likewise for the encoder $p(\boldsymbol{\xi}|\boldsymbol{s})$. One application of the VAE is to address a limitation of autoencoders, which is that their resulting latent space could have a very complex structure, i.e. small changes in latent variables could produce large variations in the data, and/or the PDF of the latent variables for an ensemble of data could be highly multimodal. By virtue of their design, VAEs tend to produce better behaved latent space PDFs. Specifically, the β-VAE is an implementation which explicitly aims to disentangle the latent space manifold via a tunable parameter β (Higgins et al. 2017), which sets the relative weighting of two competing loss functions. The first loss term is a standard reconstruction loss (mean squared error; MSE) typically adopted in autoencoders, whereas the second is the Kullback–Leibler (KL) divergence between the conditional encoder distribution $p(\boldsymbol{\xi}|\boldsymbol{s})$ and an isotropic diagonal Gaussian distribution with unit variances.

We implemented a β-VAE in the `pytorch` machine-learning framework. The same scaler transformation was applied to the training data as was used for the GPLVM, and the MSE loss was generalized to include our estimates for the continuum fitting 'error' (see equation 18).

The design of the autoencoder architecture is as follows. The encoder constitutes an MLP with two hidden layers. The first layer is non-linear, and maps the input spectral pixels to 1024 neurons passing through a SELU activation. Then two distinct linear layers convert these 1024 outputs into the mean, $\boldsymbol{\mu}$, and log variance, $\ln \boldsymbol{\sigma}^2$ (both have size n_{latent}) of the normal distribution which forms the probalistic description of the latent variables underlying the VAE. During training or evaluation, a sample from this normal distribution produces a latent space vector that is then passed through the decoder. Our decoder constitutes an MLP with two linear layers, one that maps the latent space vector to 1024 outputs, and another that maps these 1024 outputs to the n_λ data vector.

Following standard practice, we optimized to determine the MLP weights via stochastic gradient descent using the Adam optimizer with weight decay. We adopted a 90–10 per cent split of the 14 781 training spectra, using 90 per cent for training and 10 per cent for validation, where the validation set was used to evaluate performance on data unseen during training and mitigate against overfitting. The loss was computed from a mini-batch size of $N_{\text{batch}} = 128$ training set quasars per optimizer step, or epoch. To prevent overfitting, the validation set loss was also computed each epoch, and the best model was chosen to be the one that achieved the smallest value of the validation set loss. A learning rate of 10^{-4} was used and we employed early stopping, which is to say that we stopped training if the validation loss did not improve after 100 epochs. Hyperparameters were tuned via trial and error and with a more rigorous grid scan strategy. Surprisingly, we found that a very small value of $\beta = 10^{-9}$ produced the best results, which reduces the influence of the KL loss

term. Nevertheless, visual inspection of the latent space indicated that it was not significantly multimodal.

3.3 Comparison of DR algorithms

In this subsection, we compare and contrast the performance of the DR algorithms discussed in the previous section. Since our ultimate goal is to fit models to quasar spectra, the quantitative metric that we adopt for evaluating DR algorithm performance will be an estimate of the variance of the relative reconstruction error $\boldsymbol{\delta}$ defined in equation (2). How does one determine the $s_{\text{DR}}(\boldsymbol{\eta})$ representation of the spectrum $\boldsymbol{s}$ which appears there? Whereas for PCA and autoencoders, the inverse function $\boldsymbol{\xi}(\boldsymbol{s})$ is tractable and easy to evaluate, the same is not true for GPLVMs, for which determining the encoding of a new test data point requires additional assumptions (Lawrence 2005). Since our specific application of DR is the construction of a parametric model to fit quasar spectra, the most natural definition of $\boldsymbol{\eta}$ is to fit the function $s_{\text{DR}}(\boldsymbol{\eta})$ to the spectrum $\boldsymbol{s}$. That is, for each spectrum in our test set, we determine the value $\boldsymbol{\eta}_{\text{true}}$ that minimizes the mean square error

$$\text{MSE} = \sum_\lambda \left(s_\lambda - s_{\text{DR},\lambda}(\boldsymbol{\eta}) \right)^2, \tag{23}$$

where the 'true' subscript denotes that this represents the best DR representation of the quasar continuum for the hypothetical case where we directly fit a noiseless spectrum with no IGM absorption. Note that the sum in equation (23) weights each spectral pixel equally in the MSE loss computation. Indeed, the formally correct choice for the relative weighting of the terms in the equation above is rather subtle[11] and here we adopt a uniform weighting for simplicity.

As previously discussed, the hyperparameter governing the dimensionality of the latent space, n_{latent}, sets the number of free parameters $n_{\text{DR}} = n_{\text{latent}} + 1$, and generally the variance of $\boldsymbol{\delta}$ (equation 2) will decrease with increasing n_{latent} because the DR model is more flexible. We trained each of the aforementioned DR algorithms on our training set data for n_{latent} from 1 to 25 in unit steps from 1–15 and then in steps of two from 15–25. After training, we fit the models to both the 778 spectra in our autofit test set (see Section 3.1.1) and the 121 hand-fit continua in our hand-fit test set (see Section 3.1.2). Fig. 2 shows examples of our fits from the PCA, GPLVM, and β-VAE algorithms for $n_{\text{latent}} = 5$ for five randomly selected quasars from the autofit test set, where the lower set of panels shows $\boldsymbol{\delta}$.

We define a simple summary statistic for the purposes of comparing the different DR algorithms. The covariance, $\boldsymbol{\Delta}$, of the relative reconstruction error (see equation 4) is defined by

$$\Delta_{\lambda\lambda'} = \langle (\delta_\lambda - \langle\delta\rangle_\lambda)(\delta_{\lambda'} - \langle\delta\rangle_{\lambda'}) \rangle, \tag{24}$$

where the angle brackets denote an average over the members of the test set. The variance of $\boldsymbol{\delta}$ as a function of wavelength is given by the

---

[11]Our formalism in Section 2 suggests that one should take the covariant relative reconstruction errors into account when fitting for $\boldsymbol{\eta}$ (see equation 4), which would amount to maximizing a likelihood implied by equation (4) which would differ markedly from the uniform weighting in the MSE in equation (23). There is thus a chicken-and-egg problem in that, formally the likelihood one should maximize to fit for $\boldsymbol{\eta}$ depends on the mean, $\langle\delta\rangle$, and covariance, $\boldsymbol{\Delta}$, of the relative reconstruction error (equation 4), which can only be determined by analysing the distribution of residuals $\boldsymbol{\delta}$ of an ensemble of such fits. For this reason, we adopt the simple approach of fitting for $\boldsymbol{\eta}$ with uniform unit 'errors' as in equation (23) and define $\boldsymbol{\delta}$ to be the relative reconstruction error resulting from those fits.
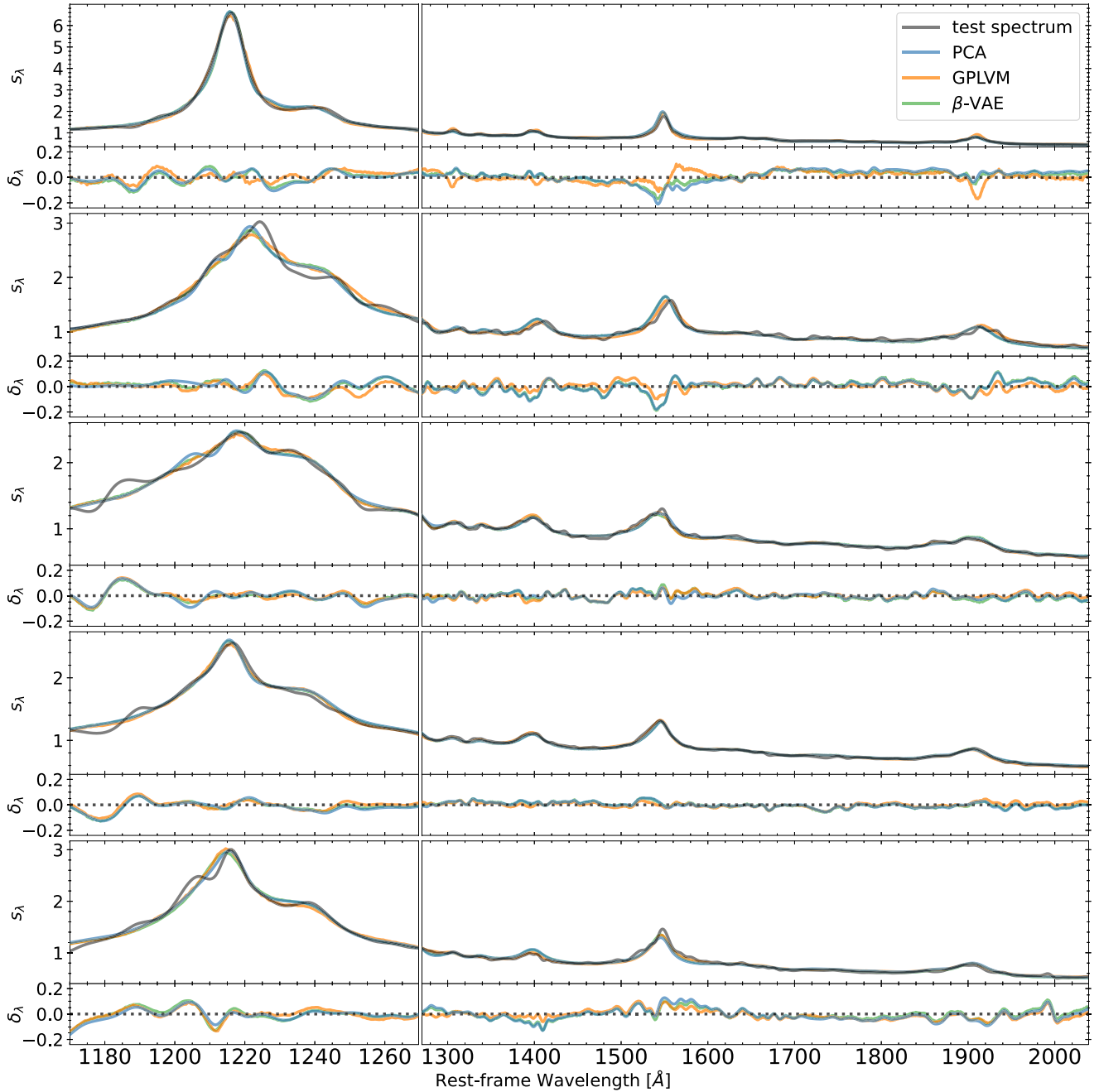
Figure 2. DR reconstructions of five randomly selected quasars from our the SDSS test set for $n_{\text{latent}} = 5$. Upper panels show the quasar continuum s_λ where the black curves are the autofit continua and the blue, yellow, and green curves are the DR reconstructions for the PCA, GPLVM, and β-VAE, respectively. The lower panels show the relative reconstruction errors, δ_λ, defined in equation (2), where the horizontal dotted line indicates the zero level.

diagonal elements of the covariance matrix $\Delta_{\lambda\lambda}$. As our metric for comparing DR algorithms, we adopt the root-mean-square variation per spectral pixel with the mean computed over the N_λ spectral pixels in the wavelength interval $[\lambda_{\text{min}}, \lambda_{\text{max}}]$

$$\text{RMS}(\lambda_{\text{min}}, \lambda_{\text{max}}) \equiv \left(\frac{1}{N_\lambda} \sum_{\lambda_{\text{min}}}^{\lambda_{\text{max}}} \Delta_{\lambda\lambda} \right)^{1/2}. \qquad (25)$$

We consider a blue region $[\lambda_{\text{min}}, \lambda_{\text{max}}] = [1185\,\text{Å}, 1260\,\text{Å}]$ covering the continuum relevant to the proximity zone and IGM damping

wing, and a red region $[\lambda_{\text{min}}, \lambda_{\text{max}}] = [1260\,\text{Å}, 2000\,\text{Å}]$ constituting the rest of the spectrum.

Fig. 3 shows DR algorithm performance as a function of n_{latent} for the autofit test set (left) and the hand-fit test set (right) for the blue wavelength range, and Fig. 4 is the analogous plot for the red wavelengths. These plots generalize the canonical explained variance versus number of components plots that one constructs in applications of PCA that are used to determine which dimensionality to compress down to. All of the DR algorithms show the expected trend of decreasing $\text{RMS}(\lambda_{\text{min}}, \lambda_{\text{max}})$ with increasing n_{latent} − smaller
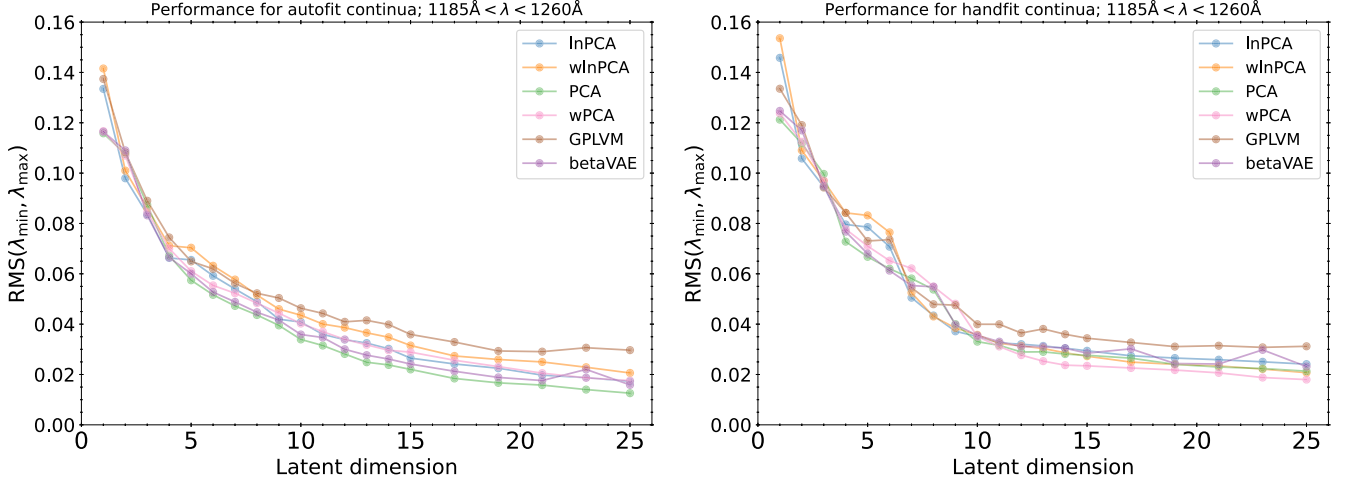
Figure 3. Comparison of DR algorithm performance for 'blue' pixels in the wavelength range 1185 Å $< \lambda <$ 1260 Å for the six DR algorithms considered. The curves show the summary statistic RMS(λ_{min}, λ_{max}) (see equation 25), for the relative reconstruction error, δ_λ (see equation 2 or the lower panels of the examples in Fig. 2) as a function of latent dimension. Left panels show results when the DR algorithms were fit to the 778 autofit continua in the SDSS test data set, whereas the right panel shows the same for the 121 hand-fit continua. Comparison of the left and right panels indicates that the behaviour of the DR algorithms does not depend on the test set used. Furthermore, the fact that the level of relative reconstruction error as a function of latent dimension is very nearly the same for autofit and hand-fit continua, suggests that the automated algorithms do not add a significant amount of extra 'noise' to the continua.
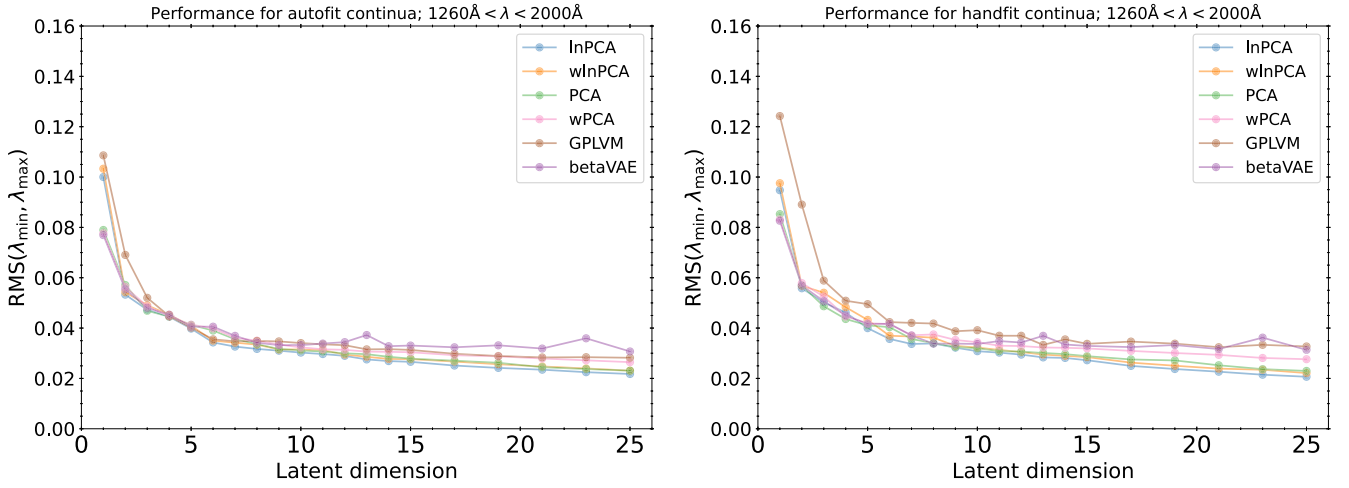
Figure 4. Comparison of DR algorithm performance for 'red' pixels in the rest-frame wavelength range 1260 Å $< \lambda <$ 2000 Å for the six DR algorithms considered. Curves and panels are the same as Fig. 3. The difference in amplitude as compared to Fig. 3 likely occurs because the ratio of pixels in emission lines to pixels in the continuum is smaller in the 1260 Å $< \lambda <$ 2000 Å wavelength range, as compared to 1185 Å $< \lambda <$ 1260 Å, where the Lyα λ1216 Å, N V λ1240 Å, and Si II λ1260 Å emission lines constitute a large fraction of the pixels and inflate the variation.

reconstruction errors will always result from a more flexible model. The most striking conclusion that one draws from Figs 3 and 4 is that simple linear DR, i.e. standard PCA or lnPCA, performs as well or better than the more sophisticated ML based non-linear DR techniques. Furthermore, this conclusion holds for both the autofit (left panels) and hand-fit (right panels) test sets, and the comparable values of RMS(λ_{min}, λ_{max}) achieved indicates that both continuum test data sets have comparable levels of noise. Finally, there appears to be no obvious advantage to the variations of PCA that we discussed, standard PCA performs just as well as lnPCA or weighted PCA. These results motivate us to adopt PCA as our DR algorithm, and to use the larger autofit test set continua for the construction of mock data in the rest of this work.

How should we choose the value of the n_{latent} hyperparameter? As is often the case with DR algorithms, the rate of decrease of the reconstruction error for a unit increment of n_{latent} varies with n_{latent}. All curves decrease steeply around n_{latent} of a few, and all flatten out at the largest values of n_{latent}. Such behaviour is expected on physical grounds – the steep decrease occurs as one approaches the number of parameters required to capture the salient features of quasar spectra, whereas the flattening occurs when this number is significantly exceeded. At the largest values of n_{latent}, incremental decreases in reconstruction error likely result from fitting noise in the autofit or hand-fit continua, or 'intrinsic noise' in the quasar continuum stochastic process, marginally better with a more flexible model. Often in the DR literature, one sets the n_{latent} hyperparameter

to the location of a 'knee' in such plots where the slope begins to flatten. Whereas for red wavelengths, Fig. 4 shows the reconstruction error starts to saturate around 2.5–3 per cent around $n_{\text{latent}} \simeq 6$, for blue wavelengths the curves in Fig. 3 exhibit a more gradual trend with n_{latent} with 2.5–3 per cent errors reached around $n_{\text{latent}} \simeq 10$. It is possible that the red wavelengths compress harder because the approximately power-law continuum of quasar spectra contributes relatively more pixels to this wavelength range, whereas the blue wavelengths are dominated by strong emission lines complexes from Lyα 1216 Å, N v 1240 Å, and S II 1260 Å that exhibit more intrinsic variation in shape. Another possibility is that IGM absorption blueward of Lyα results in more noise in the continuum estimates, making it harder to describe and compress the blue wavelengths. Our companion paper quantifies the variation in the precision with which the two IGM damping wing parameters (volume-averaged IGM neutral fraction $\langle x_{\text{HI}} \rangle$ and the quasar lifetime t_{Q}) can be recovered as a function of n_{latent} (Kist et al. 2025). There, adopting the same autofit continuum test data set used here for the mock spectra (see Section 4.2), it is found that parameter precision does not increase significantly beyond $n_{\text{latent}} > 5$, which motivates us to use $n_{\text{latent}} = 5$ as our fiducial value for the rest of this paper, corresponding to $n_{\text{DR}} = 6$ after adding the additional normalization parameter, s_{norm}.

3.4 Rebinning onto coarser wavelength grids

As discussed in Section 3.1.1, our DR test and training set spectra cover from $1170 - 2040$ Å with $dv = 140$ km s^{-1} pixels resulting in $n_\lambda = 1190$ total spectral pixels. However, these continua are smooth and can be interpolated onto a finer grid to construct mock observations of quasars with a finer spectral sampling. Whereas a spectrum of a high-z quasar might have fixed spectral resolution (given by the instrument full width at half-maximum, FWHM in km s^{-1}) and spectral sampling (the number of pixels per FWHM), for the purposes of IGM damping wing analysis, there are several reasons why rebinning the spectrum onto a coarser wavelength grid is advantageous. Note that here rebinning refers not to interpolation, but rather to averaging the flux values of finer grid pixels that land within a given low-resolution pixel. First, at a typical resolution of FWHM = 100 km s^{-1} a uniform velocity grid would result in $\gtrsim 10^3$ pixels, making the matrix computations in the likelihood in equation (17) costly to evaluate during HMC sampling. A fine velocity grid is only required blueward of rest-frame Lyα where there are narrow Lyα forest absorption lines, whereas redward of the Lyα line, the smooth damping wing signature and the broad emission lines imply that a coarse velocity grid will suffice. Second, as we discuss further in Section 5, adopting a Gaussian form (equation 10) for the transmission PDF, $P(t|\boldsymbol{\theta})$, in quasar proximity zones is an approximation which compromises the fidelity of the inference with the likelihood in equation (17). However, rebinning a high-resolution spectrum onto a coarser wavelength grid, ameliorates the problems with the inference, because it follows from the central limit theorem that the averaging involved in rebinning Gaussianizes the non-Gaussian Lyα forest transmission stochastic process, making the Gaussian approximation more valid.

Motivated by these considerations, one can consider 'hybrid' wavelength grids which have a distinct velocity grid with size dv_{blue}, for blue wavelengths $\lambda < \lambda_{\text{blue-red}}$, where it may be necessary to resolve narrow absorption features in the proximity zone, concatenated with a coarse uniform grid with pixel size dv_{red} for red wavelengths $\lambda > \lambda_{\text{blue-red}}$ for which the damping wing absorption and intrinsic quasar spectrum are smooth. Whereas such hybrid grids are considered in our companion paper (Kist et al. 2025) to assess the

impact of spectral resolution on the resulting measurement precision, in this paper we simply adopt $dv = dv_{\text{blue}} = dv_{\text{red}} = 500$ km s^{-1} throughout.

3.5 Properties of the PCA decomposition

Having arrived at PCA with $n_{\text{latent}} = 5$ as our preferred DR algorithm we now briefly describe and quantify the properties of the decomposition. Fig. 5 shows the mean spectrum $\langle s_\lambda \rangle$ and the $n_{\text{latent}} = 5$ PCA basis spectra. Recall that our approach differs from previous work using PCA to predict quasar continua (e.g. Suzuki et al. 2005; Pâris et al. 2011; Eilers et al. 2017a; Davies et al. 2018c; Ďurovčíková et al. 2020, 2024; Bosman et al. 2021) in that we do not construct separate red- and blue-side PCA decompositions, but rather a single PCA decomposition of the entire spectral range (the blue–red split in Fig. 5 is only to better visualize the blue side). PCA basis vectors are sorted in terms of the amount of variance they explain, and we see that the first component A_1 describes correlated variation in the strengths of various broad emission lines, with the subsequent components encoding more subtle correlated variations in shape.

We now compute the mean, $\langle \boldsymbol{\delta} \rangle$, and covariance, $\boldsymbol{\Delta}$, of the relative reconstruction error (see equation 4) required to evaluate the likelihood in equation (17). In general, these must be computed from the test set data on the wavelength grid adopted (i.e. see the rebinning discussion in Section 3.4) for the IGM damping wing fits. Here, we show results for the wavelength grid used for the mock quasar spectra generated in this paper (see Section 4.2), which extends from $1185 - 2000$ Å with a uniform pixel size of $dv = 500$ km s^{-1}, resulting in $n_\lambda = 313$ pixels. To obtain the $\boldsymbol{\delta}$ field we interpolate each test set spectrum s onto this wavelength grid using a cubic spline and fit for $s_{\text{DR}}(\boldsymbol{\eta})$. As we discussed in the previous section, we fit for $\boldsymbol{\eta}_{\text{true}}$ by minimizing the MSE loss function in equation (23). Adopting the uniform weighting of the spectral pixels that we discussed there,[12] we fit the PCA with $n_{\text{DR}} = 6$ (i.e. $n_{\text{latent}} = 5$ plus the normalization parameter s_{norm}) to each member of the autofit and the hand-fit test sets.

Fig. 6 shows the mean $\langle \boldsymbol{\delta} \rangle$ (dashed curves) and standard deviation $\sigma_{\boldsymbol{\delta}} \equiv \Delta_{\lambda\lambda}^{1/2}$ (solid curves; see equation 24) of the relative reconstruction error evaluated from the autofit (red) and hand-fit (blue) test sets. For the autofit test set, we see that the fits are unbiased – the mean is consistent with zero to within ~ 0.1 per cent. The factor of $\sim 2 - 4$ larger variations in the mean for the hand-fit test set results from the factor of ~ 6 times fewer quasars and as a result possibly increased sensitivity to outliers. The results for $\sigma_{\boldsymbol{\delta}}$ are consistent between the two sets and indicate typical relative reconstruction errors of $\sim 4 - 5$ per cent, which are relatively independent of wavelength, although larger errors occur in the vicinity of the strong emission lines, particularly the C IV $\lambda 1549$ line.

Finally, we can visualize the full covariant structure of $\boldsymbol{\delta}$ by computing the correlation matrix

$$\text{Corr}(\boldsymbol{\Delta})_{\lambda\lambda'} \equiv \frac{\boldsymbol{\Delta}_{\lambda\lambda'}}{\sqrt{\boldsymbol{\Delta}_{\lambda\lambda}\boldsymbol{\Delta}_{\lambda'\lambda'}}}, \qquad (26)$$

which is shown in Fig. 7 computed from the autofit test set. One clearly sees that the fit residuals are highly correlated for

---

[12]Note we use a uniform grid with $dv = 500$ km s^{-1} in this paper so each pixel receives the same weight. However, on a hybrid wavelength grid with different blue and red pixel scales, a naive uniform weighting implies the more numerous finer blue-side pixels would receive a higher relative weight. To equally treat all spectral regions, one would then need to instead upweight the red pixels by the factor $dv_{\text{red}}/dv_{\text{blue}}$.
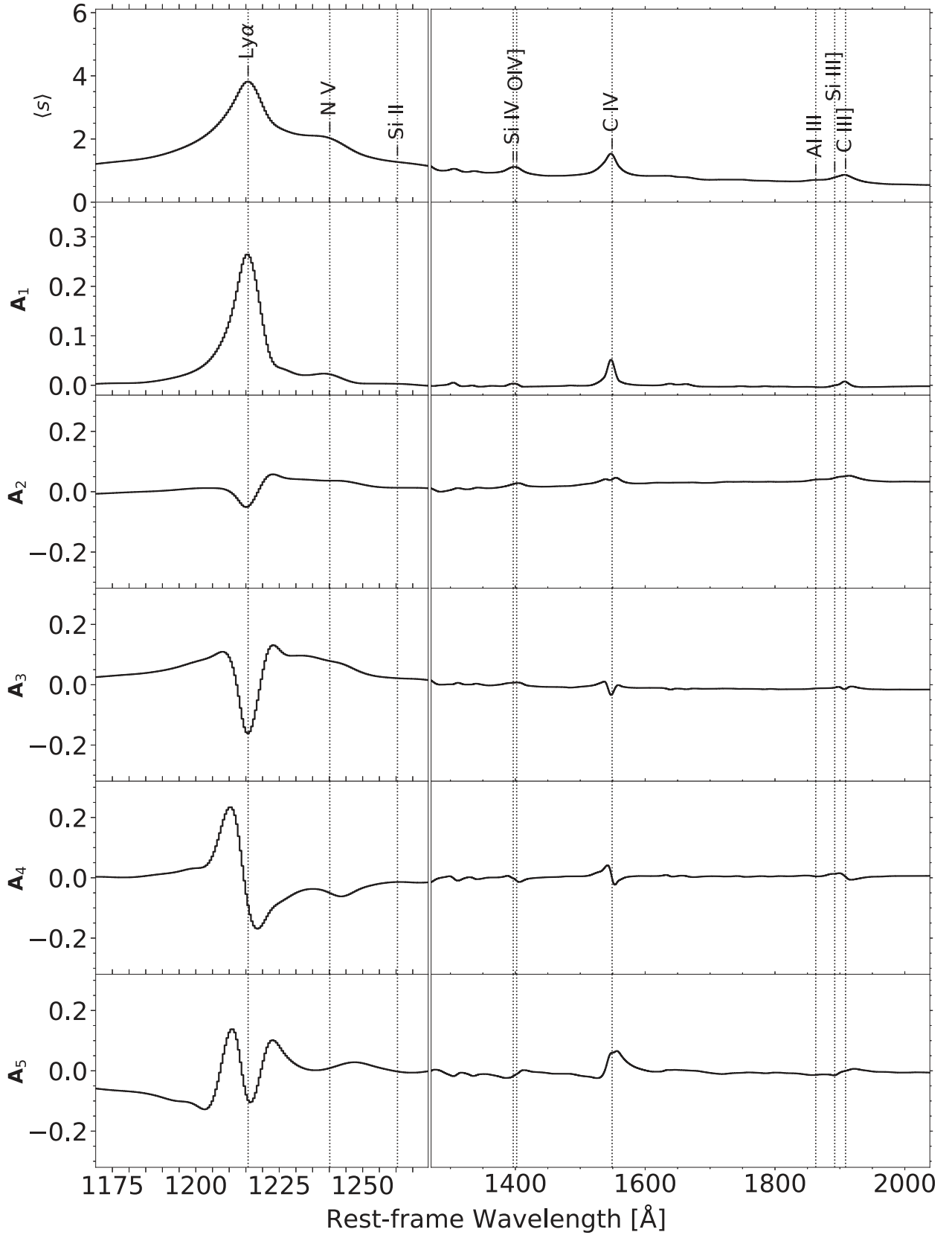
Figure 5. PCA vectors from the PCA decomposition of 14 781 SDSS quasar spectra for $n_{\text{latent}} = 5$. The upper panel labelled $\langle s \rangle$ represents the mean spectrum, while the lower panels show the principal components ordered in terms of the amount of explained variance from top to bottom. Strong broad emission lines in the quasar spectra are indicated by the vertical dashed lines and labelled in the upper panel.
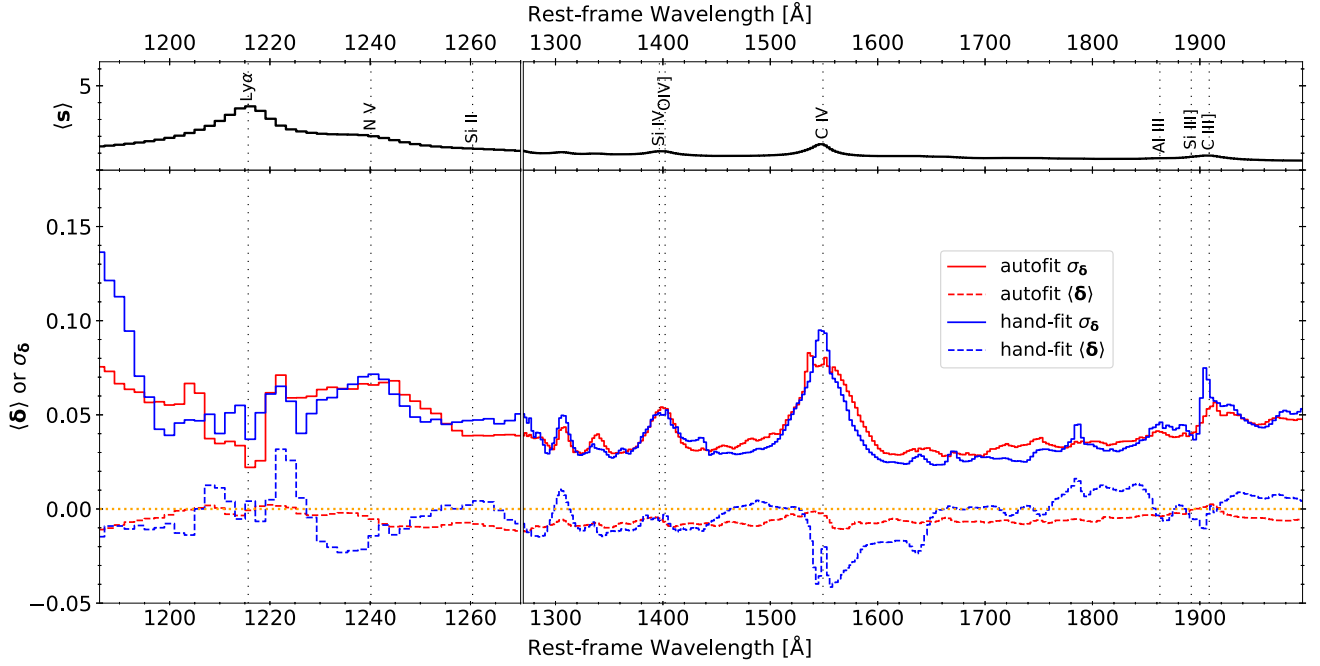
Figure 6. Moments of the relative reconstruction error as a function of wavelength for the PCA model with $n_{latent} = 5$. The lower panels show the mean $\langle \delta \rangle$ (dashed) and standard deviation $\sigma_\delta \equiv \Delta_{\lambda\lambda}^{1/2}$ (solid) of the relative reconstruction error evaluated from the 778 spectra in the autofit (red) and 121 spectra in the hand-fit (blue) test sets, respectively. For reference, the upper panel shows the mean quasar spectrum $\langle s \rangle$ constructed from the autofit test set with prominent emission lines labelled.
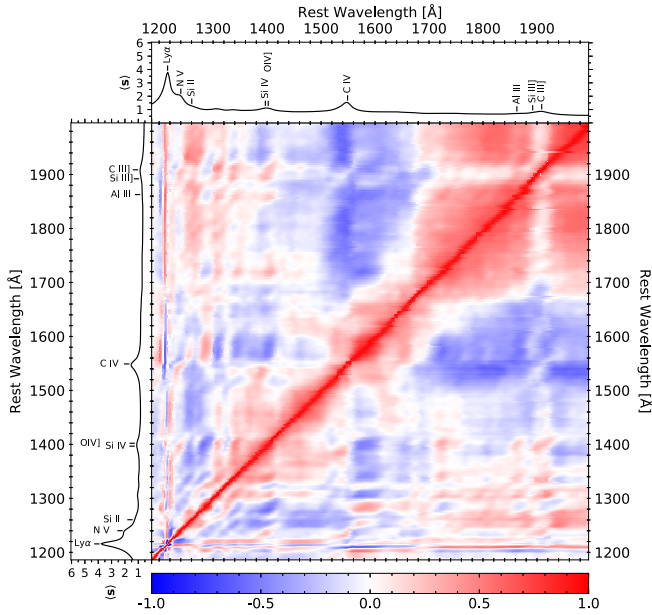
Figure 7. Correlation matrix of the relative reconstruction errors for the PCA model with $n_{latent} = 5$ evaluated from 778 autofit test set spectra. The mean quasar spectrum $\langle s \rangle$ constructed from the autofit test set is shown for reference and prominent emission lines are labelled.

neighbouring pixels as is also apparent from the lower panels showing the relative reconstruction error, δ, in Fig. 2. However, whereas correlations are typically positive in continuum regions, in the vicinity of the stronger emission lines like Lyα and C IV, one sees oscillation between correlation and anticorrelation at smaller wavelength separations. Such behaviour is expected since in general

there is more small-scale wavelength structure in the emission lines than in the smoother continuum. As for the correlations of residuals between different emission lines, the structure appears rather complex. For example, residuals around the peak of C IV are anticorrelated with the Al III–Si III]–C III] complex at ~1900 Å, whereas for pairs of pixels from the C IV and Lyα-N V emission-line complex, one observes both correlated and anticorrelated residuals.

Fig. 8 allows us to visualize the reliability of our approximation that the distribution of relative reconstruction errors is a multivariate Gaussian (see equation 3). The left and centre panels show the distribution of δ_λ for two different wavelengths relevant to studies of IGM damping wings, $\lambda = 1217.90$ Å (left) and $\lambda = 1230.35$ Å (centre) for the 778 spectra in our autofit test set. The right panel shows the distribution of the quantity

$$\chi = \mathbf{B}^{-1}(\delta - \langle \delta \rangle), \tag{27}$$

where $\mathbf{B} = \mathbf{V} \cdot \mathrm{diag}(\sqrt{\lambda})$, is the product of the matrix of eigenvectors, $\mathbf{V}$, of the covariance matrix $\mathbf{\Delta}$, and the diagonal matrix, $\mathrm{diag}(\sqrt{\lambda})$, formed from the square root of the vector of eigenvalues λ of $\mathbf{\Delta}$. For a stochastic variable described by a general multivariate Gaussian distribution (i.e. equation 3), the transformed variable χ will be a draw from $\mathcal{N}(\chi; \mathbf{0}, \mathbf{I})$ where $\mathbf{0}$ is a vector of zeros and $\mathbf{I}$ is the identity matrix. The histogram shows the distribution of $\mathbf{B}$ for all 313 spectral pixels for the 778 autofit test set spectra. Inspection of the distribution χ marginalized over the spectral pixel dimension, generalizes the intuitive method of using $(x - \mu)/\sigma$ to assess the Gaussianity of a stochastic variable x, to the case of a multivariate Gaussian distribution with a non-diagonal covariance matrix. The panels in Fig. 8 all paint a similar picture for the distribution of relative reconstruction errors. Namely, the distribution generally follows a Gaussian shape, but has stronger tails. These strong tails result in larger variance estimates than implied by the values in the 'core' of the distribution.
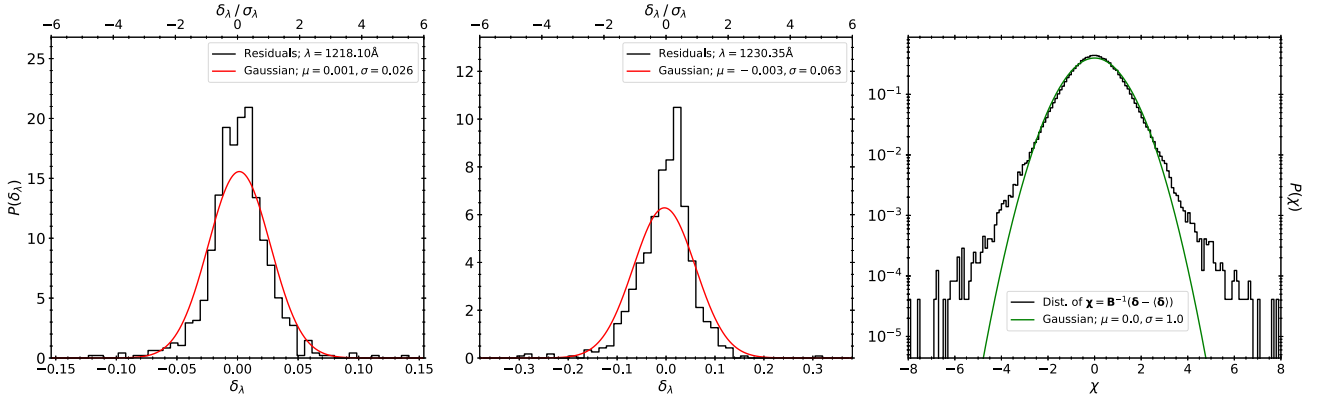
Figure 8. Distribution of relative reconstruction errors determined from fits to 778 autofit test set spectra. Left and centre: histograms show the distribution of δ_λ for two different wavelengths relevant to studies of IGM damping wings, $\lambda = 1217.90$ Å (left) and $\lambda = 1230.35$ Å (centre). Red curves show a normal distribution $\mathcal{N}(\delta_\lambda|\mu, \sigma)$ with mean and standard deviation set to that measured from the distribution. The histograms appear consistent with Gaussianity but with stronger tails. Right: the histogram shows the distribution of transformed residuals χ (see equation 27) for the full set of spectral pixels and quasars. The green curve shows a zero-centred unit variance normal distribution $\mathcal{N}(\chi|0, 1)$ for comparison. If the relative reconstruction errors $\boldsymbol{\delta}$ follow a multivariate Gaussian distribution (equation 3), then the distribution of χ should follow the green curve. Similar to the left and centre panels, the histogram appears consistent with Gaussianity but with stronger tails.

4 SIMULATING DAMPING WING OBSERVATIONS

This section describes our procedure for generating mock $z > 7$ quasar spectra with IGM damping wing absorption. First, we introduce the IGM damping wing simulations used to generate transmission skewers, then describe how these skewers are combined with our autofit continuum test set and a noise model to create realistic mock spectra. While we adopt a specific physical model to generate the reionization topology and use a hydrodynamical simulation to describe the IGM opacity, we note that our approach does not explore uncertainties in the underlying reionization model or the impact of galaxy formation physics on the IGM opacity. We will discuss these limitations further when we summarize our conclusions in Section 6.

4.1 Damping wing simulations

We generate IGM transmission skewers following the procedure developed in Davies et al. (2018b) to simulate spectra of quasar proximity zones with IGM damping wing absorption and quasar lifetime effects. Here, we briefly summarize the most important elements and refer the reader to Davies et al. (2018b) for additional details. Owing to the complexity of the problem, this is a hybrid model that combines three main components: (1) a high-resolution density field from a large volume cosmological hydrodynamical simulation, (2) a reionization topology generated from a seminumerical model of reionization, and (3) 1D ionizing RT to treat the impact of the quasar's ionizing radiation on its surrounding IGM.

Density, velocity, and temperature skewers are taken from the $z = 7.0$ output of a Nyx hydrodynamical simulation (Almgren et al. 2013; Lukić et al. 2015) in a box with side equal to 100 h^{-1}cMpc run with a 4096^3 baryon grid and 4096^3 dark matter particles. The skewers were extracted along an axis of the simulation domain and were chosen to originate on the 200 most massive dark matter haloes, $M_{\rm halo} \gtrsim 2 \times 10^{11}$ M$_\odot$, identified with a custom halo finder adapted specifically to grid hydro codes (see Sorini et al. 2018, for additional details).

To compute a realistic topology of reionization around massive quasar-hosting haloes, we use a modified version of the 21CMFAST[13] code (Mesinger, Furlanetto & Cen 2011), with an improved treatment of the ionizing photon mean-free path (Davies & Furlanetto 2022). As the Nyx simulation volume is too small to characterize the distribution of ionized/neutral regions around the rare massive haloes hosting quasars, we compute the ionization field in a larger volume, 400 cMpc on a side. The seminumerical model starts with cosmological initial conditions using the Zel'dovic approximation (Zel'Dovich 1970) generated on a 2048^3 grid, and then produces evolved density and ionization fields output at a lower resolution of 512^3. We keep all parameters that govern the reionization topology fixed except the ionizing efficiency, ζ, which sets the total number of ionizing photons emitted per collapsed baryon. Increasing ζ decreases the fraction of the volume that is neutral and vice-versa. By tuning ζ, we generate ionization fields with global volume-averaged neutral fractions $\langle x_{\rm HI}\rangle = 0.05 - 0.95$, in steps of $\Delta\langle x_{\rm HI}\rangle = 0.05$. The 21CM-FAST code constructs dark matter haloes from the initial conditions following the approach of Mesinger & Furlanetto (2007). Starting at the location of the 500 most massive haloes $M_{\rm halo} = 3 \times 10^{11}$ in the 400 cMpc domain, we randomly sampled them to extract 1200 randomly oriented skewers of the neutral fraction, $x_{\rm HI}$, for each value of the parameter $\langle x_{\rm HI}\rangle$. We trivially add a completely ionized model, $\langle x_{\rm HI}\rangle = 0.0$, and a completely neutral model, $\langle x_{\rm HI}\rangle = 1.0$, by setting $x_{\rm HI}$ to a constant everywhere to arrive at 21 total reionization models spanning the range $\langle x_{\rm HI}\rangle = 0.0 - 1.0$.

To model the impact of the quasar radiation on the IGM, we perform ionizing RT using an updated version of the 1D RT code described in Davies et al. (2016), which computes the time-dependent evolution of the ionized fractions of hydrogen and helium, as well as the gas temperature. This time evolution is governed by the quasar's radiative history, and we assume a so-called light bulb light curve parametrized by the quasar lifetime, t_Q, whereby the quasar turned on at some point t_Q in the past, and has been shining at constant luminosity ever since. The ionizing photon output from the quasar was computed using the Lusso et al. (2015) spectral energy
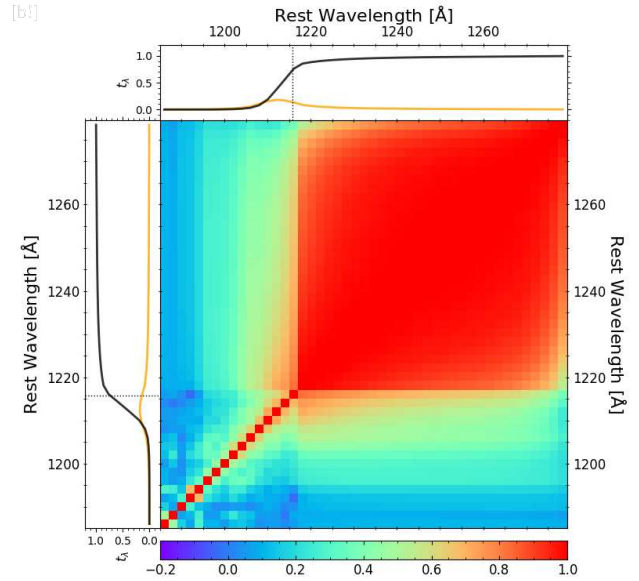
---

[13] https://github.com/andreimesinger/21cmFAST

distribution normalized to agree with the photometric measurements of the quasar. In this paper, we use models constructed for the quasar ULAS J1342+0928 at $z = 7.54$, which has an apparent J-band AB magnitude of $J_{AB} = 20.3$ (Bañados et al. 2018).

To generate IGM transmission skewers originating from quasars, we use the density and temperature fields from the Nyx simulations and the x_{HI} skewers from the seminumerical ionization topology as initial conditions for the RT code. For ionized regions which have values of $x_{HI} = 0$ in the seminumerical ionization skewers, we initialize to $x_{HI} \sim 10^{-3}$ by adopting a fixed amplitude of the UVB photoionization rate, Γ_{HI}. The IGM temperature was initialized to the values from the Nyx simulation in ionized regions, whereas we assume a temperature of $T = 2000$ K inside neutral regions. For each value of the global IGM neutral fraction $\langle x_{HI} \rangle$ parametrizing the reionization topology, we ran RT calculations using 1200 distinct random Nyx skewers (six per halo for 200 Nyx haloes), and associated them with each of the 1200 seminumerical x_{HI} skewers. The RT was computed on a uniform logarithmic grid of quasar lifetimes spanning from $\log_{10}(t_Q/\text{yr}) = 3 - 8$ in steps of $\Delta \log_{10}(t_Q/\text{yr}) = 0.1$ dex (this grid of RT computations were generated in Davies et al. 2019) and IGM transmission skewers were computed over the velocity range $-10,000 \text{ km s}^{-1} \leq \Delta v \leq 10,000 \text{ km s}^{-1}$ on a uniform velocity grid with 5000 pixels of size $dv = 4 \text{ km s}^{-1}$. This was done by performing the optical depth velocity integral using the neutral hydrogen density and temperature outputs from the RT, the peculiar velocity field from the Nyx simulation, and weighting by the Voigt line profile that describes the frequency dependence of the Lyα absorption cross-section. Because the Nyx simulations do not have star or galaxy formation subgrid recipes, hydro grid elements in collapsed structures can evolve to have unrealistically high densities, which would not be present in a simulation with galaxy formation prescriptions. These high overdensities are problematic for several reasons. First, the RT can get stuck modelling self-shielding from such high-density systems, because the convergence criteria push the adaptive time-step to very short values. Second, when analysing IGM damping wings, objects with strong proximate absorbers (PDLAs or PLLSs), are excluded from analysis, so it is sensible to do the same for the simulations. To mitigate the undesired effects of unphysically large densities, we omit the first 35 pkpc from consideration when extracting the density, temperature, and velocity fields from the Nyx simulation, and we clip the gas overdensity to always be below 200, i.e. roughly the virial overdensity. Nevertheless, even for the combination of model parameters where proximate DLAs are least expected ($x_{HI} = 0.0$ and $\log_{10}(t_Q/\text{yr}) = 8$), we still found that a small fraction (17 of 1200) of the simulated IGM transmission spectra exhibited proximate DLA absorption,[14] which we chose to exclude from consideration for all models. Thus, the final output of our IGM transmission simulations is a set of 1183 transmission spectra at each location of a coarse 21×51 grid corresponding to the two IGM parameters, $\boldsymbol{\theta} = (\langle x_{HI} \rangle, \log_{10}(t_Q/\text{yr}))$. Figures illustrating our numerical simulation procedure can be found in Davies et al. (2018b).

To evaluate the likelihood in equation (17), we require the mean, $\langle t(\boldsymbol{\theta}) \rangle$, and covariance, $\boldsymbol{C}_t(\boldsymbol{\theta})$ of the proximity zone Lyα transmission as a function of model parameters $\boldsymbol{\theta} = (\langle x_{HI} \rangle, \log_{10}(t_Q/\text{yr}))$. Fig. 9

Figure 9. Mean transmission and covariance structure of the proximity zone Lyα transmission field $\boldsymbol{t}$ evaluated from the 1183 transmission spectra for a model $\langle x_{HI} \rangle = 0.50$ and $\log_{10}(t_Q/\text{yr}) = 6$ on a velocity grid with $dv = 500 \text{ km s}^{-1}$. Top and left panels show the mean transmission $\langle \boldsymbol{t} \rangle$ (black) and its 1σ variation $\sqrt{\text{diag}(\boldsymbol{C}_t)}$ (orange), whereas the 2D image shows the correlation matrix with the level of correlation indicated by the colour bar. The wavelength of Lyα at line centre, 1215.67 Å, is shown by the vertical dotted line. The smooth coherent IGM damping wing signature results in highly correlated transmission fluctuations redward of Lyα; whereas, for $\lambda \lesssim 1215.67$ Å, the correlations are smaller owing to Lyα forest fluctuations in the proximity zone, although they are not insignificant.

illustrates these quantities at the wavelengths relevant for IGM damping wings on our velocity grid with $dv = 500 \text{ km s}^{-1}$ pixels for a model with $\langle x_{HI} \rangle = 0.50$ and $\log_{10}(t_Q/\text{yr}) = 6$. The top and left panels show the mean transmission $\langle \boldsymbol{t} \rangle$ (black) and its 1σ variation $\sqrt{\text{diag}(\boldsymbol{C}_t)}$ (orange). The covariant structure of the transmission can be visualized via the correlation matrix shown in the 2D image, computed from $\boldsymbol{C}_t$ (see the definition in equation 26). The smooth coherent IGM damping wing signature results in highly correlated transmission fluctuations redward of Lyα; whereas, for $\lambda \lesssim 1215.67$ Å, the correlations are smaller owing to Lyα forest fluctuations, although not insignificant.

4.2 Creating mock quasar spectra

In this section we describe our procedure for generating realistic mock quasar spectra with IGM damping wings. As described in Section 4.1, we focus on models of the quasar ULAS J1342+0928 at $z = 7.54$ with $J_{AB} = 20.3$. The simulated spectra cover the rest-frame (observed-frame) wavelength range $1185 - 2000$ Å ($10120 - 17080$ Å). As discussed in Section 3.4, it is advantageous to rebin the spectra from the native sampling of the simulations onto a coarser wavelength grid with $dv = 500 \text{ km s}^{-1}$. We assume observations at a spectral resolution described by a Gaussian line-spread function (LSF) with FWHM $= 100 \text{ km s}^{-1}$. To forward model the impact of spectral resolution and rebinning onto the coarse $dv = 500 \text{ km s}^{-1}$ grid, we convolve our ensemble of $1183 \times 21 \times 51$ IGM transmission skewers (1183 skewers, 21×51 models) defined on the simulation spectral grid $dv_{sims} = 4 \text{ km s}^{-1}$, with this Gaussian LSF and then rebin them onto the coarse wavelength grid. Mock

---

[14]These were identified as skewers with transmission deviating by more than 2 per cent from unity at a pixel 1000 km s^{-1} redward of the quasar redshift for the model with $x_{HI} = 0.0$ and $\log_{10}(t_Q/\text{yr}) = 8$. This model choice is conservative, since it is least likely to exhibit damping wing absorption both because the IGM is reionized and the longest quasar lifetime maximizes the likelihood of photoevaporating proximate absorbers.

quasar spectra are generated by randomly selecting a continuum from our 778 autofit test set, cubic spline interpolation of the continuum onto the coarse wavelength grid, and then multiplying by a randomly selected resolution convolved and rebinned IGM transmission skewer.[15]

We simulate observational data collected with a ground-based 8-m class telescope with a hypothetical instrument with a constant throughput of 30 per cent. To correctly model the noise, we used the SKYCALC_IPY[16] PYTHON package to generate realistic models of the sky background and atmospheric transmission, which are used to construct a full spectrum simulator including telluric absorption and noise contributions from object photons, sky background, and detector read noise (assumed to be four electrons per $d\upsilon = 500$ km s^{-1} spectral pixel). We tuned a hypothetical exposure time to achieve a median signal-to-noise ratio of S/N = 10 per 100 km s^{-1} velocity interval, computed over the telluric absorption free observed frame (rest-frame) wavelength range $11750 - 13300$ Å ($1376 - 1557$ Å). This in turns allows us compute the correct relative contributions of photon counting and detector read noise to the noise budget, resulting in a realistic noise vector $\boldsymbol{\sigma}$. Multiplying this noise vector into a random draw from a unit variance Gaussian distribution, generates a realistic realization of heteroscedastic noise (due to OH sky lines and telluric absorption features), which is then added to the mock quasar spectrum. Three examples of mock quasar spectra generated via this procedure are shown in Fig. 10.

5 INFERENCE RESULTS

In this section, we present the results from statistical inference performed on the mock EoR quasar spectra that were introduced in Section 4.2 using the new likelihood we derived in Section 2 (see equation 17). First we discuss how we use HMC to sample the posterior distribution, then we describe the coverage test we perform to assess the reliability of our statistical inference. Our inference turns out to be overconfident and thus does not pass the coverage test, but we describe a procedure that reweights the HMC samples to remedy this problem and thus perform reliable inference. After showing examples of our inference at work, we build intuition for why our inference fails the coverage test. We conclude by quantifying how well we recover the underlying quasar continuum and compare the accuracy of our reconstructions with past work.

5.1 Hamiltonian Monte Carlo

HMC (Duane et al. 1987) is a powerful variant of the traditional Markov Chain Monte Carlo (MCMC) algorithm for sampling probability distributions in high dimensions (see e.g. Betancourt 2017, for a review). Based on a powerful analogy with Hamiltonian dynamics, HMC introduces auxiliary momentum variables that interact with the target variables representing the samples from the distribution of interest. HMC numerically integrates the equations of Hamiltonian

dynamics, and then interleaves Metropolois–Hastings steps to accept or reject the proposed target states ensuring that detailed balance is satisfied. Compared to MCMC, HMC capitalizes on the gradients of the log-likelihood, making it more efficient for sampling in high-dimensional spaces and handling correlated target variables. This advantage enables HMC to generate distant proposals more effectively, leading to reduced autocorrelation in samples, better performance for multimodal distributions, and faster convergence.

The main challenge of using HMC is that it requires gradients of the log-likelihood function with respect to the model parameters, which can be challenging or costly to compute. Gradients can be computed in several ways: by numerical approximation using finite differences, through symbolic manipulation of analytic expressions, or via automatic differentiation (AD; e.g. Baydin et al. 2018), a programmatic application of the chain rule that yields exact derivatives at machine precision by tracing the operations in a computation. In our case, we rely on AD, which yields exact gradients for arbitrary compositions of differentiable operations and scales efficiently even in high-dimensional parameter spaces. Modern AD frameworks automate this process, making it straightforward to obtain derivatives even for complicated likelihood functions involving matrix operations, linear algebra, and complex control flow. The introduction of AD environments in PYTHON such as JAX (Bradbury et al. 2018), makes exploiting AD to compute the required gradients for HMC straightforward.

We developed PYTHON software to compute an automatically differentiable version of the likelihood in equation (17) in JAX. For the HMC, we use the No-U-Turn Sampler (NUTS; Hoffman & Gelman 2014) variant of HMC implemented in the JAX-based NUMPYRO (Bingham et al. 2018; Phan, Pradhan & Jankowiak 2019) probabilistic programming package in the PYTHON module `numpyro.infer.NUTS`. The NUTS sampler was run with four chains of 1000 samples, each with a warmup phase of 1000 samples, using the 'vectorized' chain option. The NUTS algorithm builds a binary tree that is used to take forward/backwards 'directional' steps to explore the target posterior using gradients to guide it towards the highest probability regions. The `max_tree_depth` parameter, which sets the size of this binary tree, was set to 10 (i.e. up to a maximum of 1024 steps for each iteration). As alluded to in Section 3.4, the run time is a strong function of the number of pixels, $n_{\rm pix}$, in the quasar spectrum, as this sets the dimensionality of the matrices in equation (17), and several of the required matrix operations scale as $\mathcal{O}(n_{\rm pix}^3)$. In general, there are order unity variations in the runtime which depend on details of the warmup phase that constructs the 'mass matrix' determining the Hamiltonian dynamics in HMC, which in turn will depend on the parameter-dependent shape of the underlying posterior distribution. The wavelength grid adopted in this paper has $n_\lambda = 313$. Running on a single thread of an AMD EPYC 7763 2.45 GHz processor,[17] the typical runtimes were ~ 2.5 hours. Running on a single thread of an Intel Xeon Gold 6126 2.6 GHz processor[18] equipped with a NVIDIA GeForce RTX 2080TI GPU (with 11 GB of memory), the typical runtimes were ~ 15 min.

5.2 Coverage testing

Let $\boldsymbol{\Theta}$ be the parameter vector of interest for a Bayesian inference problem that will be applied to experimental data $\boldsymbol{x}$. In the current context of IGM damping wings, $\boldsymbol{\Theta} = \{\boldsymbol{\theta}, \boldsymbol{\eta}\}$, where $\boldsymbol{\theta}$ are the

---

[15]Note that the two operations, (1) multiplication of the full-resolution IGM transmission with the continuum, and (2) convolution of the resulting spectrum with the LSF and rebinning do not formally commute. The correct order is to multiply by the continuum first and then convolve and rebin, but the differences are negligible. We choose to convolve and rebin the IGM transmission skewers first, and then multiply by the continuum afterward because the expensive convolution and rebinning operations can then be performed only once in pre-processing, dramatically speeding up all downstream computations.

[16]https://github.com/AstarVienna/skycalc_ipy

[17]These processors have 64 cores, 128 threads, and 256 MB of L3 Cache.

[18]These processors have 12 cores, 24 threads, and 19.25 MB of L3 Cache.
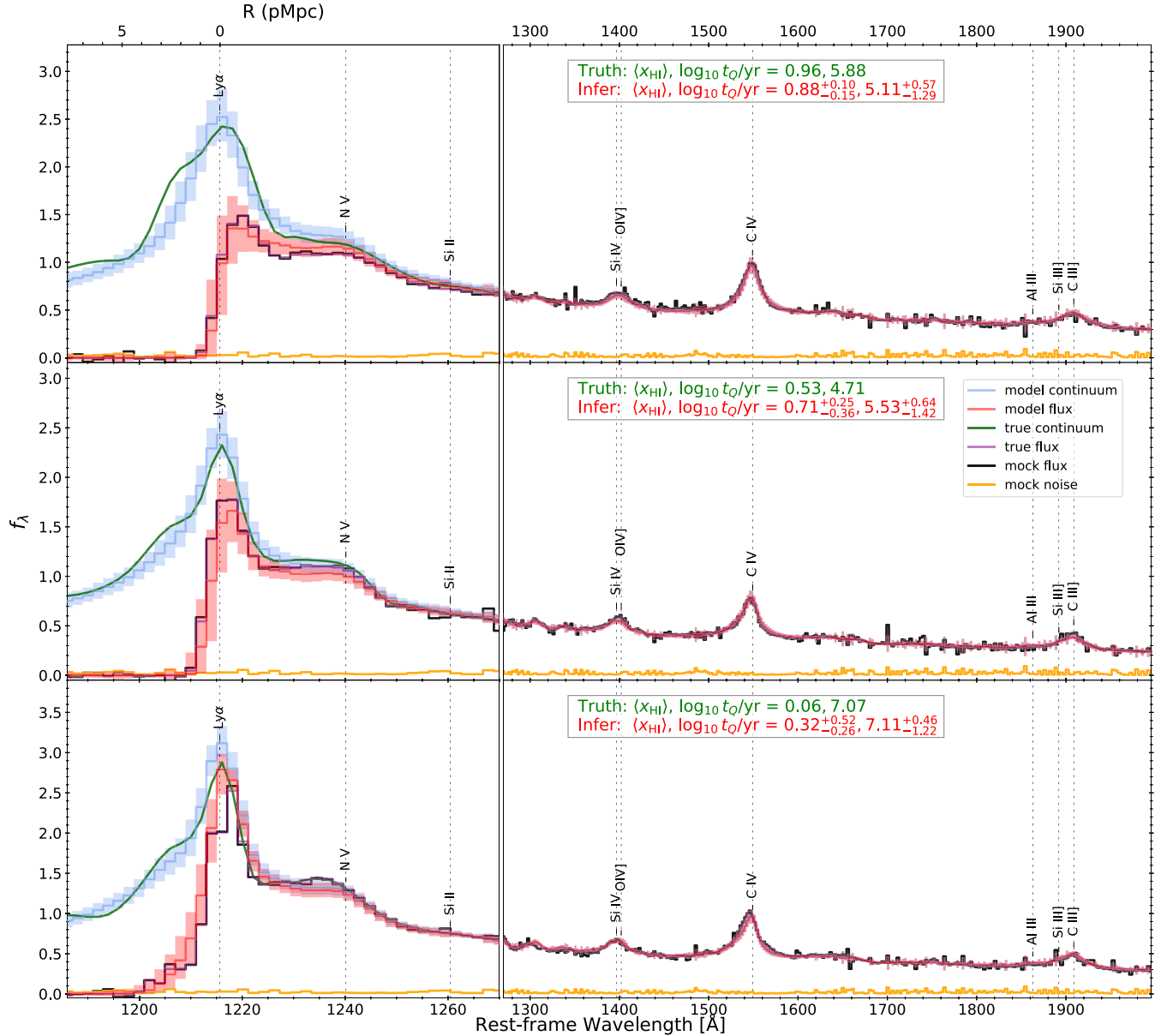
Figure 10. Examples of inference performed on mock spectra for three different combinations of $\langle x_{HI} \rangle$ and $\log_{10}(t_Q/\text{yr})$. Black shows the noisy mock quasar spectra, orange the 1σ spectral noise level, purple the true noiseless flux, green the true underlying continuum, blue the median inferred continuum model, and red the median inferred damping wing model flux. Shaded regions indicate the 16th and 84th percentiles of the continuum model (blue) and model flux (red). For the continuum model, the median and shaded regions incorporate parameter variations, continuum reconstruction errors, and spectral noise, whereas for the model flux the incorporate these effects as well as IGM transmission fluctuations (see Section 5.5 for details). True model parameters are annotated in green, whereas median inferred parameters and 16th and 84th percentile credibility intervals (determined from marginalized posteriors) are annotated in red. The inferred constraints are those determined after the marginal coverage correction weights (which slightly dilate the contours) have been applied (see Section 5.4). A corner plot illustrating the full 8D posterior for the mock with $\langle x_{HI} \rangle = 0.96$ and $\log_{10}(t_Q/\text{yr}) = 5.88$ in the upper panel is shown in Fig. 12 (also coverage corrected).

astrophysical parameters and $\boldsymbol{\eta}$ are the DR parameters, and the data $\boldsymbol{x}$ are the quasar spectrum $\boldsymbol{f}$ and its associated noise vector $\boldsymbol{\sigma}$. Following Bayes theorem, the posterior distribution for $\boldsymbol{\Theta}$, given experimental data $\boldsymbol{x}$ is

$$P(\boldsymbol{\Theta}|\boldsymbol{x}) = \frac{L(\boldsymbol{x}|\boldsymbol{\Theta})P(\boldsymbol{\Theta})}{P(\boldsymbol{x})}, \quad (28)$$

where $L(\boldsymbol{x}|\boldsymbol{\Theta})$ is the likelihood of the data given the model, $P(\boldsymbol{\Theta})$ reflects our prior knowledge of the parameters, and $P(\boldsymbol{x})$ is known as the evidence, which is interpreted as a normalization constant since

it is independent of the parameters $\boldsymbol{\Theta}$. If the inference procedure is reliable, one expects that the probability α obtained by integrating the posterior probability density in equation (28) over a volume V of parameter space $\boldsymbol{\Theta}$,

$$\alpha = \int_{V_\alpha} P(\boldsymbol{\Theta}|\boldsymbol{x})\mathrm{d}\boldsymbol{\Theta} \quad (29)$$

corresponds to a true probability. In other words, that the parameter space volume enclosed by the 68 per cent credibility contour contains the true parameters 68 per cent of the time under repetitions of
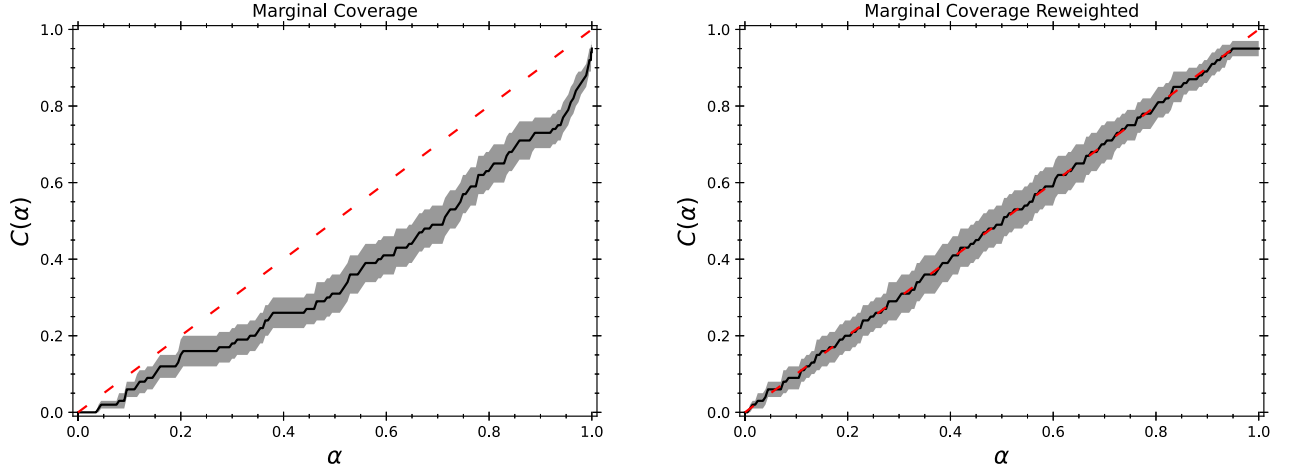
Figure 11. Coverage test results determined from $N = 100$ mock spectra. Left: the black line shows the coverage $C(\alpha)$ of the marginal posterior distributions $P(\boldsymbol{\theta}|\boldsymbol{f}_j, \boldsymbol{\sigma}_j)$, which empirically quantifies how often the true astrophysical parameters $\boldsymbol{\theta}_{\text{true}} = (\langle x_{\text{HI}}\rangle, \log_{10}(t_{\text{Q}}/\text{yr}))$ lie within the αth credibility contour. The red-dashed line shows the expected $y = x$ curve that one would obtain for a perfect inference. The grey-shaded region shows the effective 1σ error range on $C(\alpha)$ determined from 16th and 84th percentile ranges of the binomial distribution. Given that $C(\alpha)$ lies everywhere below the red-dashed line, we see that our pipeline systematically delivers overconfident inference. Right: coverage test results after enlarging the credibility contours by reweighting the HMC samples. By construction, $C(\alpha) = \alpha$, indicating that on average the inference pipeline delivers reliable credibility contours.

the experiment, and analogously for all the other credibility levels. However, this need not always be the case. Imperfections in an inference procedure can cause the posterior distribution to exclude the true parameters more or less often than indicated by the nominal credibility contours, because the posterior is shifted, or too narrow and hence overconfident, or too broad and hence underconfident, or combinations thereof. Such imperfections can result from bugs in an inference pipeline, or from an inappropriate choice of the prior, or by adopting an approximate likelihood that fails to accurately capture the statistical properties of the measurement process. In the current context, we know that the likelihood in equation (17) is only approximate, as discussed in Section 2, which motivates us to explore its coverage.

The coverage probability, $C(\alpha)$, of a posterior credibility level α is the fraction of the time that the true parameters lie within the volume enclosed by the corresponding credibility contour under repetitions of the experiment (see e.g. Sellentin & Starck 2019). It provides a mathematically rigorous method to quantify whether a posterior distribution delivers reliable probabilities. If an inference procedure is robust, then the coverage probability should equal to the posterior credibility for every level. This approach of comparing coverage probabilities to posterior credibility levels is referred to as an 'inference test' or a 'coverage test'.

We conduct a coverage test following the algorithm described in more detail in Appendix A (see also Wolfson et al. 2023), which we now summarize:

(i) Draw $N = 100$ astrophysical parameter vectors $\boldsymbol{\theta}_{\text{true},j}$ from uniform priors defined by $\langle x_{\text{HI}}\rangle \sim \text{Uniform}(0, 1)$ and $\log_{10}(t_{\text{Q}}/\text{yr}) \sim \text{Uniform}(3, 8)$. These are the 'true' parameters that generate the mock data sets used to perform the coverage test.

(ii) Generate realizations of mock quasar spectra $\{\boldsymbol{f}_j, \boldsymbol{\sigma}_j\}$ from these 'true' parameters following the approach described in Section 4.2.

(iii) Perform HMC inference on each data set as described in Section 5.1 resulting in a set of 2000 samples for the astrophysical parameters $\boldsymbol{\theta}$ and continuum nuisance parameters $\boldsymbol{\eta}$, from each of the $N = 100$ posterior distributions $P(\boldsymbol{\theta}, \boldsymbol{\eta}|\boldsymbol{f}_j, \boldsymbol{\sigma}_j)$.

(iv) Consider a set of M credibility contour levels $\alpha \in [0, 1]$. For each value α and each mock, test whether the true astrophysical parameter values, $\boldsymbol{\theta}_{\text{true},j}$, reside within the volume V_α enclosed by αth contour. For each α, the coverage probability $C(\alpha)$ is the fraction of the N mock data sets for which the true values lie within the volume V_α defined by equation (29).

As described in Appendix A, a coverage test can be performed for the entire parameter vector, here $\boldsymbol{\Theta} = \{\boldsymbol{\theta}, \boldsymbol{\eta}\}$ using the full posterior $P(\boldsymbol{\theta}, \boldsymbol{\eta}|\boldsymbol{f}_j, \boldsymbol{\sigma}_j)$, as well as for the astrophysical (i.e. non-nuisance) parameters $\boldsymbol{\theta}$ using their marginal posterior $P(\boldsymbol{\theta}|\boldsymbol{f}_j, \boldsymbol{\sigma}_j)$. Obviously, the coverage of the marginal posterior $P(\boldsymbol{\theta}|\boldsymbol{f}_j, \boldsymbol{\sigma}_j)$ is most critical, since nuisance parameters will be marginalized out, which is what we focus on in what follows.

For α, we consider a vector of 200 linearly spaced values in the range $\alpha \in [0, 0.994]$ concatenated with a vector of 101 linearly spaced values in the range $\alpha \in [0.995, 1.0]$, resulting in $M = 301$ values of α. As will be apparent in the next section, the steep $C(\alpha)$ versus α curve near $\alpha \sim 1$ (see Fig. 11) motivates adopting a more finely spaced grid as α approaches unity.

5.3 Marginal coverage test results

In Fig. 11, we show the marginal coverage test results determined from the approach described in Section 5.2 and Appendix A for $N = 100$ mock quasar spectra. The black line shows the coverage $C(\alpha)$ of the marginal posterior distributions $P(\boldsymbol{\theta}|\boldsymbol{f}_j, \boldsymbol{\sigma}_j)$, which empirically quantifies how often the true astrophysical parameters $\boldsymbol{\theta}_{\text{true}} = (\langle x_{\text{HI}}\rangle, \log_{10}(t_{\text{Q}}/\text{yr}))$ lie within the αth credibility contour. The red-dashed line shows the expected $y = x$ curve that one would obtain for a perfect inference pipeline in the limit $N \to \infty$. An overconfident inference pipeline will yield a curve $C(\alpha)$ versus α that lies systematically below the line $y = x$, whereas for an underconfident inference procedure the $C(\alpha)$ will lie above $y = x$. As discussed in Appendix A (see also Sellentin & Starck 2019), since $C(\alpha)$ counts how often the true parameters fall inside the αth contour, it is the number of successes in a sequence of N independent experiments each asking a yes-or-no question. Hence, $C(\alpha)$ follows

the binomial distribution $B(N, C(\alpha))$, which we use to assign errors. The grey-shaded region shows the effective 1σ error range[19] on $C(\alpha)$ determined from 16th and 84th percentile ranges of $B(N, C(\alpha))$.

Given that $C(\alpha)$ lies everywhere below the red-dashed line, we see that our pipeline systematically delivers overconfident inference. For example, for the effective 1σ credibility level $\alpha = 0.68$, i.e. the 68th percentile credibility contour, the coverage is actually just $C(0.68) = 0.48$, indicating that on average the contours include the true $\theta_{\text{true}} = (\langle x_{\text{HI}} \rangle, \log_{10}(t_Q/\text{yr}))$ only 48 per cent of the time. Similarly, for the effective 2σ credibility level of $\alpha = 0.95$ the coverage is $C(0.95) = 0.79$.

5.4 Reweighting samples to pass a coverage test

The coverage test for the marginal posterior $P(\theta | f, \sigma)$ distribution shown in left panel of Fig. 11 indicates that $C(\alpha)$ lies everywhere below the red-dashed line $y = x$, indicating that our pipeline systematically delivers overconfident inference. How can we nevertheless perform statistically reliable inference in light of this overconfidence?

In Appendix A, we introduce a novel procedure whereby the HMC posterior samples can be assigned weights, such that the reweighted samples produce reliable inference which passes a coverage test by construction (see also Wolfson et al. 2023). The mathematics underlying this procedure is described in detail in Appendix A, but it can be understood heuristically as follows. Sellentin & Starck (2019) advocate that one simply relabel the credibility contours to reflect the fact that inference is overconfident. In other words, since the coverage plot in Fig. 11 indicates that the 68th percentile contour only contains the true value 48 per cent of the time, we will simply refer to this contour as the 48th percentile rather than the 68th. The real 68th percentile contour containing the true parameters 68 per cent of the time under the inference test actually corresponds to the value $\alpha = 0.85 = C^{-1}(0.68)$ contour for our original approximate inference. In other words, we can systematically expand all of the credibility contours of the original inference by the right amount, such that they contain the true model the empirically correct fraction of the time. In general, as we show in Appendix A, this remapping of the credibility levels α into true coverage probabilities $C(\alpha)$ can be achieved by solving for the set of weights for the HMC samples from the original posterior, which by construction guarantees that we will pass a coverage test.

The purpose of HMC (or MCMC) samples from a posterior is to estimate credibility intervals on parameters, perform marginalization integrals, and compute 'moments' of the posterior via Monte Carlo integration. If we can determine the set of weights that corrects the imperfect inference such that it passes a coverage test, these weights can then be used in all of the downstream computations that one performs with the samples, guaranteeing the reliability of our statistical inference.

The right panel of Fig. 11 shows the coverage of the marginal posterior $P(\theta | f, \sigma)$ distribution after the HMC samples have been reweighted. Samples from the core of the distribution with higher $P(\theta | x)$ are downweighted, whereas samples in the outskirts of the distribution with lower $P(\theta | x)$ are upweighted, such that the net effect is to grow the credibility contours. The agreement of $C(\alpha)$ with the red-dashed $y = x$ line indicates that we now achieve perfect coverage, which as explained in Appendix A, occurs by construction because we solve for the set of weights that guarantees this outcome.

[19]Note however that the different values of α are clearly correlated since the same ensemble was used to calculate all of them.

Note that although we compute the weights to guarantee that the reweighted HMC astrophysical parameter samples, θ_j, will pass a coverage test for their marginal posterior distribution, we nevertheless apply these weights to the entire parameter vector $\Theta = \{\theta, \eta\}$. In other words, the contours for all parameters including the nuisance parameters will also be dilated. In practice, this implies that we pass the coverage test perfectly for the astrophysical parameters (by construction, as shown in the left panel of Fig. 11), θ, whereas our inference is slightly underconfident for the full parameter vector, Θ, which includes the nuisance parameters, η (i.e. the contours are slightly too large). Since underconfident inference is more conservative, and as this applies only to the nuisance parameters, we have demonstrated that our full pipeline passes a coverage test and delivers reliable statistical inference.

5.5 Inference examples

Examples of our inference applied to mock spectra (see Section 4.2) with reweighted HMC samples (see Section 5.4) are shown for three different combinations of $\langle x_{\text{HI}} \rangle$ and $\log_{10}(t_Q/\text{yr})$ in Fig. 10. The black histograms show the noisy mock quasar spectra, orange the 1σ spectral noise levels, green the true underlying continua, blue the median inferred continuum models, and red the median inferred damping wing model flux profiles. True model parameters are annotated in green, whereas median inferred parameters and 16th and 84th percentile credibility intervals (determined from marginalized posteriors after the coverage correction reweighting) are annotated in red.

Our approach for visualizing the quality of the fits, and hence the solid lines and shaded regions in Fig. 10, warrants further discussion. A realization of the model of the spectrum can be determined by evaluating the model of the spectra (e.g. the DR quasar continuum, the proximity zone transmission profile, or the product of the two) at each HMC sample, which has an associated weight. As discussed in Section 5.4, median model curves and model confidence intervals can be determined by computing weighted percentiles of these model curves. However, according to the likelihood in equation (17), there are multiple sources of scatter that govern how well a model curve will fit the data, namely spectral noise, IGM fluctuations, and continuum reconstruction errors. A naive visual data-model comparison that does not take all these sources of stochasticity into account can be misleading. As a concrete example, if one compares the median and 1σ interval of the PCA continnum to the quasar spectrum redward of the Lyα region, it will not appear to be a good fit relative to the spectral noise alone, since this ignores the continuum reconstruction error budget. Thus in the current context, choosing sensible 'error bars' for visual data-model comparison is rather subtle.

To generate the median and 16th and 84th percentile model intervals in Fig. 10 we proceed as follows. For each parameter vector $\Theta = \{\theta, \eta\}$ from the HMC posterior Markov chain, we:

(i) Evaluate the PCA DR model $s_{\text{DR}}(\eta)$ via equation (20).

(ii) Draw a realization of the relative continuum reconstruction error δ from the Gaussian distribution in equation (3), and compute $s = s_{\text{DR}}(\eta) \circ (1 + \delta)$.

(iii) Draw a random IGM transmission skewer t for the value of θ from the set of simulated skewers, allowing us to compute $f = s \circ t$.

(iv) Draw a realization of Gaussian spectral noise $\tilde{\sigma}$ consistent with the noise vector σ and compute $\tilde{s} = s + \tilde{\sigma}$ and $\tilde{f} = f + \tilde{\sigma}$ which we refer to as the noisy model continuum and the noisy model flux.

The blue histogram and shaded region in Fig. 10 represent the weighted median and 16th and 84th weighted percentiles of the noisy model continuum, $\tilde{s}$, respectively, which reflects parameter uncertainty, continuum reconstruction errors, and spectral noise. The red histogram and shaded region are the weighted median and the same percentiles of the noisy model flux, $\tilde{f}$, reflecting parameter uncertainty, continuum reconstruction errors, IGM transmission fluctuations, and spectral noise. Note that IGM transmission fluctuations are intrinsically accounted for by drawing IGM transmission skewers from the model for a choice of astrophysical parameters θ, whereas we are explicitly adding the continuum reconstruction errors and spectral noise by drawing realizations from their respective Gaussian distributions.

Fig. 12 shows a corner plot illustrating the full 8D posterior for the mock spectrum in the top panel of Fig. 10, which has true model parameters $\langle x_{HI} \rangle = 0.96$ and $\log_{10}(t_Q/\text{yr}) = 5.88$. The red square and the horizontal and vertical red lines indicate the true parameter values. Note that Fig. 12 shows the posterior distribution after applying the marginal coverage correction weights described in Section 5.4. Fig. 13 compares the nuisance parameter marginalized posterior distributions for the two astrophysical parameters, $\langle x_{HI} \rangle$ and $\log_{10}(t_Q/\text{yr})$ for the same mock spectrum in the top panel of Fig. 10 (i.e. the upper left panel of the full posterior in Fig. 12), before (original HMC; green) and after (reweighted; black) applying the coverage correction weights to the samples from the HMC chain (see Section 5.4). It is evident that this reweighting broadens the posterior, correcting for the overconfidence of the original marginal posterior distribution (see Fig. 11), as discussed in Section 5.4. Qualitatively, the shape of the posterior in the $\langle x_{HI} \rangle - \log_{10}(t_Q/\text{yr})$ plane resembles the shape of the posteriors recovered in the analysis of real $z > 7$ quasar spectra by Davies et al. (2018b). Specifically, the well-known degeneracy (see e.g. Bolton et al. 2011; Davies et al. 2018b) between IGM neutral fraction and quasar lifetime in determining the shape of the proximity zone and IGM damping wing profile is apparent.

5.6 Understanding the poor coverage

The overconfidence of our inference arises from the approximate form of the analytic likelihood adopted in equation (17). Specifically, the problematic approximation was the assumption of a Gaussian form for the Lyα forest transmission PDF, $P(t|\theta)$, in equation (10). We have explicitly demonstrated this by generating mocks where the simulated t transmission skewers are replaced with samples from the multivariate normal distribution in equation (10). We carried out a coverage test on $N = 100$ such Gaussianized mocks for the same set of mock quasar continua, and find that $C(\alpha) = \alpha$ within the binomially distributed counting errors, indicating that the assumption of a Gaussian form for $P(t|\theta)$ is responsible for the failure of the coverage test.

To better visualize how this non-Gaussianity manifests, one could compare the actual distribution of t from the simulated proximity zone skewers to the Gaussian adopted in equation (10). However, this would not be the entire story since the observable is not the noiseless transmission t, but rather the flux f. Notwithstanding the non-Gaussianity of $P(t|\theta)$, the resulting distribution of f could still be close to the Gaussian adopted in equation (17), because convolution with the Gaussian spectral noise (the Σ term in equation 17) and the very nearly Gaussian (see right panel of Fig. 8) relative reconstruction errors (the $\langle T \rangle C_s \langle T \rangle$ term in equation 17) could nevertheless Gaussianize the distribution of f.

However, it is easier to visualize the distribution of a continuum normalized quantity than the flux f, and this would allow us to

aggregate mocks with different intrinsic continua s on a single plot. Motivated by this, we consider the quantity $\hat{t} \equiv f \circ \langle s(\eta_{\text{true}}) \rangle^{-1}$, which we will refer to as the pseudo transmission. It is akin to the real transmission t, but the flux is divided by $\langle s(\eta_{\text{true}}) \rangle$ instead of the true continuum. Here, η_{true} is the latent variable obtained by fitting the DR model to the mock continuum with no noise or IGM absorption as discussed in Section 3.3. Recalling that $\langle s(\eta) \rangle = s_{DR}(\eta) \circ (1 + \langle \delta \rangle)$, we see that $\hat{t}$ is the flux normalized by the product of our best-fitting DR estimate to the continuum, $s_{DR}(\eta_{\text{true}})$, and the average bias $(1 + \langle \delta \rangle)$ of these estimates. Thus, the distribution of $\hat{t}$ will include scatter arising from both spectral noise and the imperfection of the DR continuum model. We can recast the likelihood in equation (17) as a PDF for $\hat{t}$, which gives

$$P\left(\hat{t}|\sigma, \theta, \eta\right)$$
$$= \mathcal{N}\left(\hat{t}; \langle t \rangle, \langle S \rangle^{-1} \Sigma \langle S \rangle^{-1} + C_t + \frac{\langle T \rangle}{1 + \langle \delta \rangle} \Delta \frac{\langle T \rangle}{1 + \langle \delta \rangle}\right), \quad (30)$$

where we recall that $\langle S \rangle \equiv \text{diag}(\langle s(\eta_{\text{true}}) \rangle)$ and hence $\langle S \rangle^{-1} \equiv \text{diag}(\langle 1/s(\eta_{\text{true}}) \rangle)$, and where we define $\frac{\langle T \rangle}{1+\langle \delta \rangle} \equiv \text{diag}\left(\frac{\langle t \rangle}{1+\langle \delta \rangle}\right)$, where as before the division of one vector by another is performed element-wise.

To further simplify the expression in equation (30), we assume the spectral noise vector can be written as $\sigma = \langle s(\eta_{\text{true}}) \rangle / \text{snr}$, where snr is a vector of spectral signal-to-noise ratio values and again division of two vectors is performed element wise. By choosing the noise to be proportional to the continuum level $\langle s(\eta_{\text{true}}) \rangle$, the dependence on $\langle S \rangle$ cancels out of the $\langle S \rangle^{-1} \Sigma \langle S \rangle^{-1}$ term of the covariance in equation (30), yielding an expression that is independent of s and η

$$P\left(\hat{t}|\sigma, \theta\right) = \mathcal{N}\left(\hat{t}; \langle t \rangle, \text{SNR}^{-2} + C_t + \frac{\langle T \rangle}{1 + \langle \delta \rangle} \Delta \frac{\langle T \rangle}{1 + \langle \delta \rangle}\right), \quad (31)$$

where $\text{SNR}^{-2} \equiv \text{diag}(1/\text{snr}^2)$. This expression allows us to express the distribution of $\hat{t}$ generated from many different mock spectra (with different s) via a single PDF, which can be easily visualized. The form of the PDF in equation (31) is intuitive – as expected $\hat{t}$ is distributed about the mean IGM transmission $\langle t \rangle$, and the total covariance is a sum of three matrices, the first SNR^{-2} quantifying spectral noise, the second C_t quantifying fluctuations of the IGM transmission field t, and the third quantifying continuum reconstruction errors.

To generate realizations of $\hat{t}$ to compare to the Gaussian distribution in equation (31), we start by generating mock IGM damping wing spectra f using the procedure described in Section 4.2 for a model with $\langle x_{HI} \rangle = 0.50$ and $\log_{10}(t_Q/\text{yr}) = 6$. The continuum S/N of these mocks is s/σ (where σ is the noise from our simulator described in Section 4.2). We then set snr to be the median value (i.e. median taken over the quasar dimension), thus adopting a single but representative value of the spectral S/N as a function of wavelength. We then regenerated the mocks with the noise $\sigma = \langle s(\eta_{\text{true}}) \rangle / \text{snr}$, with snr set to this median value, and divided the fluxes f by $\langle s(\eta_{\text{true}}) \rangle$, yielding realizations of $\hat{t}$. Whereas throughout this work we restricted to the 778 test set autofit continua for constructing mock IGM damping wing spectra, here for the sole purpose of visualization, we instead use the larger training set of 14 781 autofit continua to increase the number of samples.[20]

---

[20]As we only have 1183 IGM transmission skewers for each parameter value θ, we randomly assigned the autofit continua to an IGM transmission skewer with replacement.
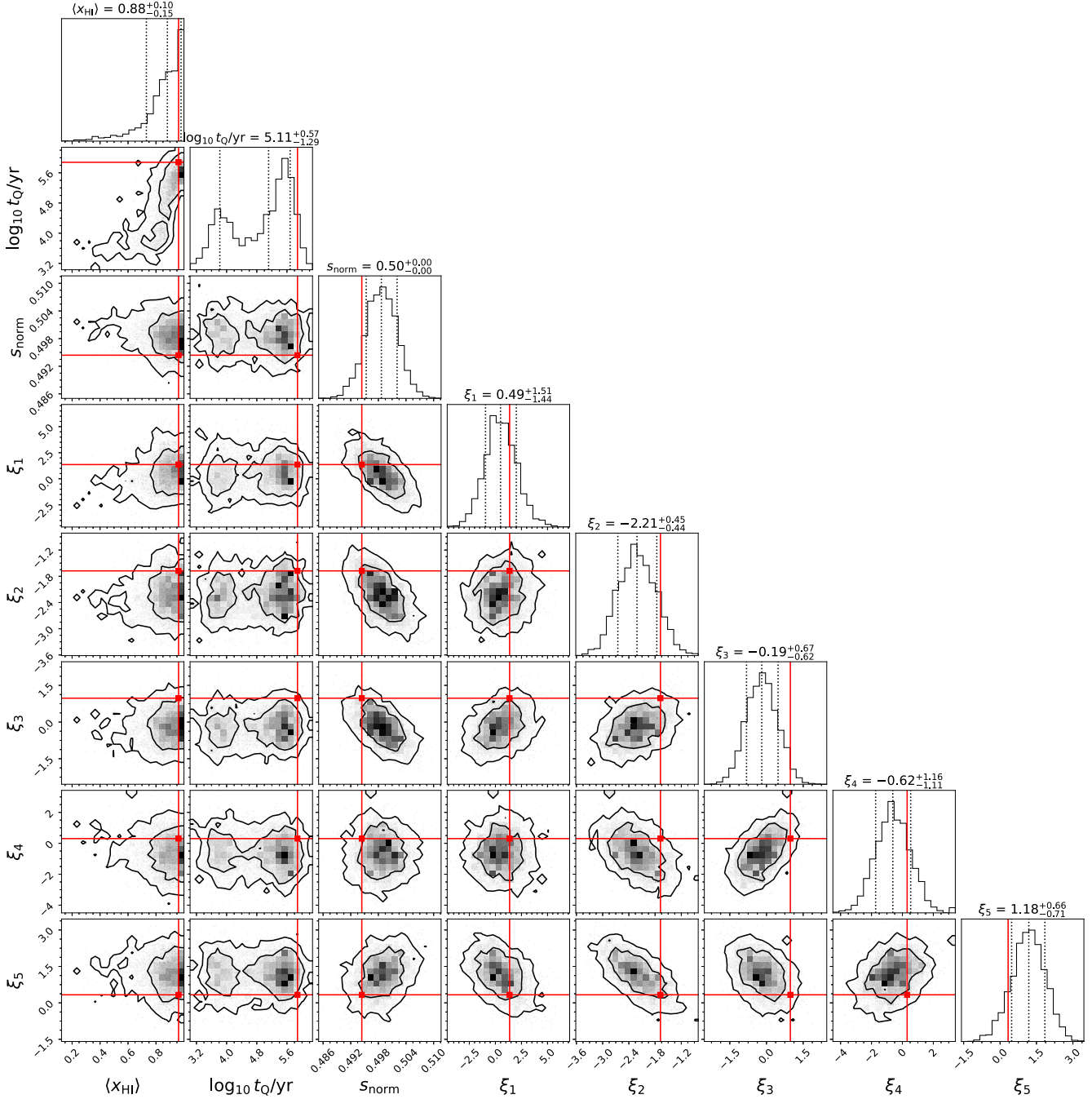
Figure 12. Corner plot illustrating the full 8D posterior distribution resulting from our statistical inference procedure applied to the mock quasar spectrum in the top panel of Fig. 10 with $x_{HI} = 0.96$ and $\log_{10}(t_Q/yr) = 5.88$. The red square and the horizontal and vertical red lines indicate the true parameter values. The posterior distribution shown here is the result after the marginal coverage correction weights described in Section 5.4 have been applied, which slightly dilates the contours relative to the original HMC inference.

To determine how well the PDF in equation (31) approximates the true distribution of $\hat{t}$, we plot slices of this distribution in the 2D $\hat{t}_\lambda - \hat{t}_{\lambda'}$ plane, where λ and λ' are the wavelengths of two different spectral pixels. Examples for two distinct pairs of transmission values are shown in Fig. 14. The left panel of Fig. 14 indicates that at rest-frame wavelengths where the mean proximity zone transmission is high, $\langle t_\lambda \rangle \sim 0.5 - 0.9$, the Gaussian approximation is decent, as indicated by: (1) the similarity of the 2D probability density contours for the $\hat{t}_\lambda - \hat{t}_{\lambda'}$ samples (solid lines) and the contours of the approximate analytical Gaussian PDF (dotted lines; equation 31),

(2) the same comparison for the 1D marginal distribution (i.e. grey marginal histograms compared to black dotted lines for the Gaussian PDF), and (3) the fact that the percentage of samples (indicated in the legend) within the 68 per cent (71.2 per cent; dotted blue) and 95 per cent (95.4 per cent; dotted green) analytical Gaussian probability density contours are very close to the expected values for each contour.

However, the right panel of Fig. 14 indicates that for bluer wavelengths where the mean proximity zone transmission is lower, $\langle t_\lambda \rangle \sim 0.1 - 0.2$, the Gaussian 2D contours and 1D marginal PDFs
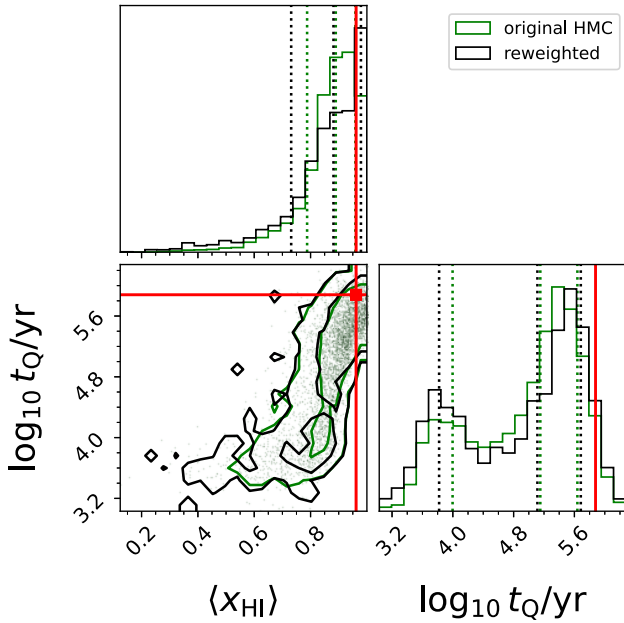
Figure 13. Comparison of the 2D marginal posterior distribution in the astrophysical parameters, $\langle x_{\mathrm{HI}} \rangle$ and $\log_{10}(t_{\mathrm{Q}}/\mathrm{yr})$ (i.e. the upper left panel of the full posterior in Fig. 12), before (original HMC; green) and after (reweighted; black) applying the coverage correction weights to the samples from the HMC chain (see Section 5.4). The red square and the horizontal and vertical red lines indicate the true parameter values. Vertical dotted lines in the marginal posterior panels indicate the 16–50–84 percentile ranges. It is evident that reweighting the samples broadens the posterior, correcting for the overconfidence (see Fig. 11) of the original marginal posterior distribution as discussed in Section 5.4.

poorly approximate the distribution of the $\hat{t}_\lambda - \hat{t}_{\lambda'}$ samples. Whereas the Gaussian PDF is symmetric about the mean $\langle t_\lambda \rangle$, the samples are significantly skewed to positive transmission values. This disagreement is easily understood – whereas a Gaussian centred at a low $\langle t_\lambda \rangle \sim 0.1 - 0.2$ must be symmetric by construction and thus predicts a significant probability for negative values, the transmission is an inherently positive quantity. Besides the positivity constraint, additional skewness of the t_λ field PDF results from the reionization topology and density fluctuations in the IGM. At intermediate and low values of the average IGM neutral fraction, $\langle x_{\mathrm{HI}} \rangle \lesssim 0.5$, the distribution of distances from quasar host haloes to the first patch of neutral gas exhibits a strong tail to large distances (see fig. 2 of Davies et al. 2018c). The corresponding reionized regions near the quasar will be further photoionized by the quasar's radiation resulting in transmissive proximity zone regions. Thus tails in the distance to the nearest neutral patch, combined with the strong tails in the in Lyα forest transmission PDF at the relevant optical depths (see e.g. Davies et al. 2018a), manifest as a tail towards high transmission, t_λ, for bluer proximity zone wavelengths where the average profile has $\langle t_\lambda \rangle \sim 0.1 - 0.2$. These heavy positive tails bias the Gaussian transmission covariance, C_t, high, relative to the width of the core of the distribution of samples (see marginal histograms in the right panel of Fig. 14). As a result the percentage of samples within the 68 per cent Gaussian contour (79.8 per cent; dotted blue) deviates significantly from the Gaussian expectation.

We conclude that despite the Gaussianizing effects of Gaussian spectral noise and approximately Gaussian continuum reconstruction errors, our approximate Gaussian form for the likelihood in

equation (17) is nevertheless a poor approximation at low average transmission values $\langle t_\lambda \rangle \sim 0.1 - 0.2$ because of the strong underlying non-Gaussianity of the IGM transmission PDF $P(t|\theta)$.

5.7 Continuum reconstruction recovery and comparison to previous work

The main advantage of the statistical inference method at the heart of this paper is that it constructs a generative model for the entire quasar spectrum, including absorbed pixels in the proximity zone, to perform parameter inference. This contrasts with the red–blue prediction approach that has been adopted in most past work (e.g. Davies et al. 2018b; Ďurovčíková et al. 2020; Fathivavsari 2020; Reiman 2020; Chen et al. 2022; Greig et al. 2024b) modelling IGM damping wings, whereby only pixels redward of some cut-off (typically $\lambda > 1280\,\text{Å}$) are used to predict the blue-side ($\lambda < 1280\,\text{Å}$) continuum and its error, which are then used to perform inference. Our approach is generically expected to perform better, given that it uses all of the information available, and specifically the spectral pixels $\lambda < 1280\,\text{Å}$ where much of the information about the intrinsic quasar continuum shape around Lyα, and all of the information about the IGM damping wing are manifest. However, quantifying the improvement in precision on the astrophysical parameters $\langle x_{\mathrm{HI}} \rangle$ and $\log_{10}(t_{\mathrm{Q}}/\mathrm{yr})$ and directly comparing to previous work would be challenging, given the heterogeneity of the different modelling approaches that have been employed to date. Furthermore, such a comparison might not be fair given that the current study is the only one to investigate (and explicitly correct for) the coverage of the inference: if the inference pipelines used in past work were overconfident, a comparison to our results would not be a fair one.

In our companion paper (Kist et al. 2025), we quantify the precision on the astrophysical parameters yielded by our inference procedure in detail by analysing an ensemble of 1000 mock spectra spanning the full astrophysical parameter range ($0.0 \leq \langle x_{\mathrm{HI}} \rangle \leq 1.0$; $3 \leq \log_{10}(t_{\mathrm{Q}}/\mathrm{yr}) \leq 8$). We find that (see table 1 and fig. 9 of Kist et al. 2025) even for the hypothetical case of perfect knowledge of the quasar continuum, the median $\sim 1\sigma$ precision on $\langle x_{\mathrm{HI}} \rangle$ is ~ 15 per cent and the median precision on $\log_{10}(t_{\mathrm{Q}}/\mathrm{yr})$ is ~ 0.55 dex. For inference on full mocks generated according to the procedure in Section 4.2, where we model and marginalize over the unknown quasar continuum, the typical measurement error increases to ~ 28 per cent on $\langle x_{\mathrm{HI}} \rangle$ and ~ 0.80 dex for $\log_{10}(t_{\mathrm{Q}}/\mathrm{yr})$. This analysis suggests that continuum reconstruction errors contribute roughly an equal amount to the error budget as the other sources of stochasticity in the problem, namely the distribution of distances to the nearest patch of neutral hydrogen (see fig. 2 of Davies et al. 2018c) and fluctuations in the location of the quasar ionization front due to the distribution of sinks along the sightline. Thus significant gains in astrophysical parameter precision should result from an algorithm that better reconstructs the quasar continuum.

Motivated by this, we perform a careful comparison of the precision of our continuum reconstructions to the PCA-based red–blue continuum prediction method introduced in Davies et al. (2018b). While this is only one of the many algorithms that have been used for quasar continuum reconstruction to date, Greig et al. (2024b) recently conducted a detailed comparison of all 10 of the quasar continuum prediction pipelines in existence, and found that while they all yield roughly comparable precision, the Davies et al. (2018b) pipeline consistently performed among the best for the various samples and metrics considered. Hence comparing the performance of our continuum reconstruction approach to the Davies et al. (2018b) algorithm
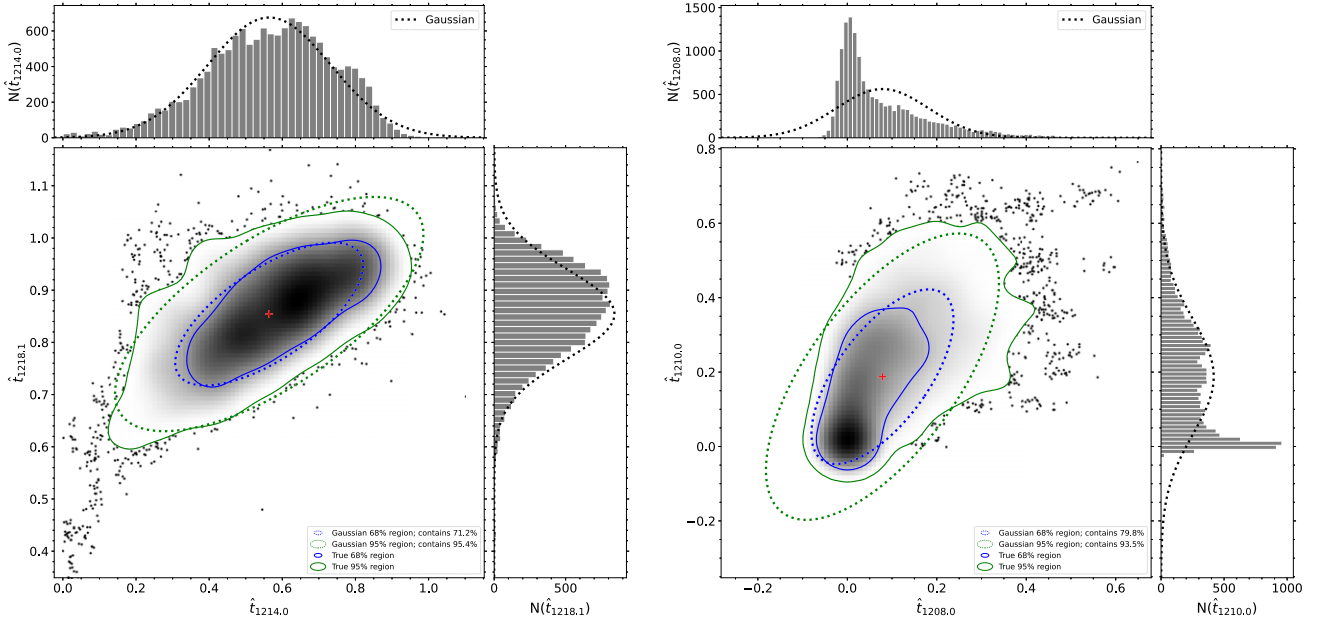
Figure 14. Left: 2D slice of the multivariate distribution of the pseudo-transmission field $\hat{t}_\lambda$ for two rest-frame wavelengths $\lambda = 1214.0$ Å (x-axis) and $\lambda = 1218.1$ Å (y-axis) in the quasar proximity zone, corresponding to mean transmission values of $\langle t_{1214.0} \rangle = 0.56$ and $\langle t_{1218.1} \rangle = 0.85$, which is indicated by the red cross. The model considered has $\langle x_{\mathrm{HI}} \rangle = 0.50$ and $\log_{10}(t_Q/\mathrm{yr}) = 6$. Shading indicates the probability density with a logarithmic stretch. Solid lines show the 68 per cent (blue) and 95 per cent (green) probability density contours determined from the distribution of $\hat{t}$ samples, whereas the dotted lines show the corresponding probability density contours for the Gaussian distribution in equation (31). The legend indicates the percentage of the samples that are within the Gaussian probability density contours. Histograms show the 1D marginal distributions where black dotted lines show the Gaussian prediction based on equation (31). Right: same as left except for a different set of rest-frame wavelengths $\lambda = 1208.0$ Å (x-axis) and $\lambda = 1210.0$ Å (y-axis), corresponding to lower values of the mean transmission $\langle t_{1208.0} \rangle = 0.08$ and $\langle t_{1210.0} \rangle = 0.19$.

should, broadly speaking, constitute a sufficient comparison to the diverse set of existing continuum prediction algorithms in existence.

To perform this comparison, we considered the $N = 100$ mock quasar spectra on the d$v = 500$ km s^{-1} wavelength grid that we conducted inference on (see Fig. 10 for examples) to arrive at the coverage test results described in Section 5.3 and shown in Fig. 11. We define the inferred quasar continuum for a given mock spectrum to be the weighted median (using the marginal coverage weights) of the PCA DR models, $s_{\mathrm{DR}}(\eta_i)$, evaluated at each of the HMC samples η_i for each mock spectrum, which we denote by $s_{\mathrm{DR, med}}$. However, the DR model itself is not perfect and will result in some relative reconstruction error, δ, even if the PCA is fit to a spectrum with no noise or IGM absorption (see Section 3.3 and Fig. 6). Thus, a proper analysis requires comparing the moments of the relative reconstruction error of the inferred continua

$$\delta_{\mathrm{inf}} = \frac{s - s_{\mathrm{DR,med}}}{s}, \tag{32}$$
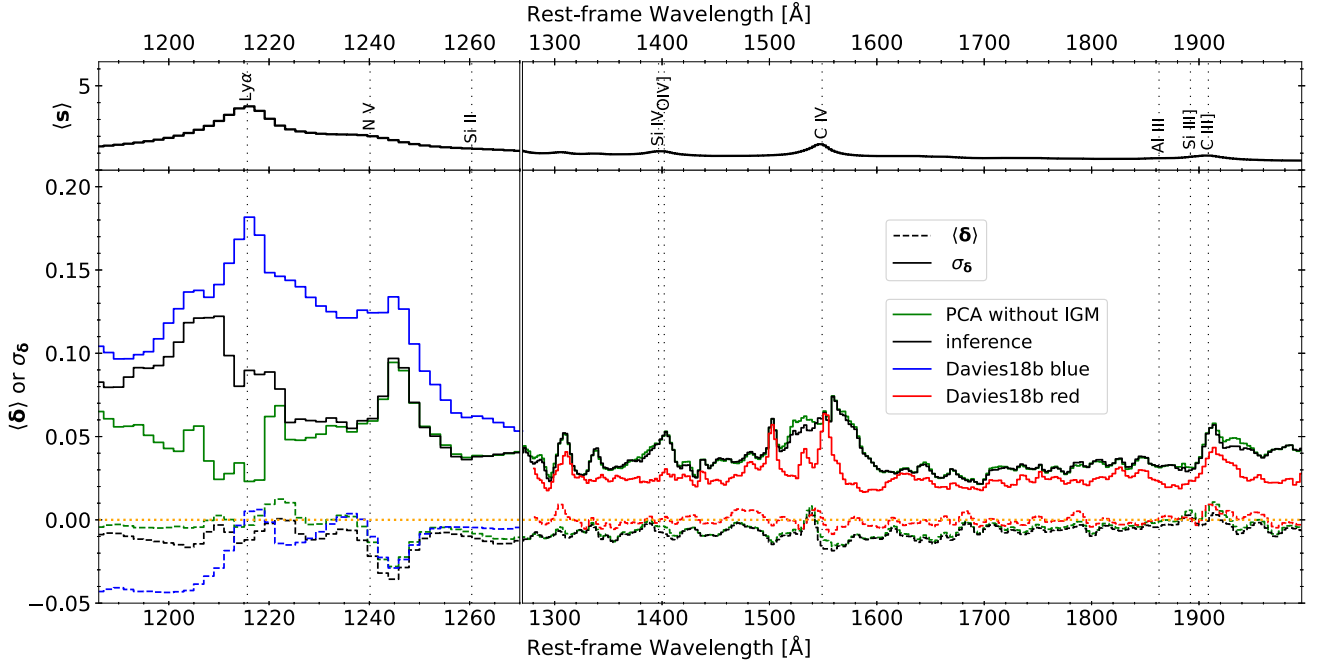
to the moments of the relative reconstruction error

$$\delta = \frac{s - s_{\mathrm{DR}}(\eta_{\mathrm{true}})}{s}, \tag{33}$$

obtained by fitting the PCA to the same mock quasar spectra, but without noise or IGM absorption. As described in Section 3.3, we determine the best-fitting PCA DR parameters, η_{true}, by minimizing the MSE loss given in equation (23), and Fig. 6 shows the moments of the relative reconstruction error, $\langle \delta \rangle$ and $\sigma_\delta \equiv \Delta_{\lambda\lambda}^{1/2}$ (see equation 24), computed from the spectra in our continuum test sets. In the current context, we will compare the moments of δ_{inf} directly to the moments of δ, both computed from the same 100 mock spectra for which we

performed statistical inference. Since the DR model always results in reconstruction errors (even in the absence of noise and IGM absorption), the best continuum reconstruction (in the presence of noise and IGM absorption) would be one for which δ_{inf} is very close to δ.

To compare the accuracy of our continuum reconstructions to those from the Davies et al. (2018b) method, we applied their red–blue prediction algorithm to the same set of 100 mock spectra. Specifically, we followed the Davies et al. (2018b) approach and fit their red-side PCA vectors to the red spectral pixels ($\lambda < 1280$ Å) of our mocks yielding 10 PCA coefficients for each mock, and then we used their transformation matrix to transform each set of 10 red-side PCA coefficients into six blue-side PCA coefficients, which finally yields a predicted blue-side continuum ($\lambda < 1280$ Å) for each mock.

The efficacy of the quasar continuum reconstruction algorithms is illustrated in Fig. 15. In the lower panel, the dashed lines show the mean $\langle \delta \rangle$ of the relative reconstruction errors, whereas the solid lines show the standard deviation $\sigma_\delta \equiv \Delta_{\lambda\lambda}^{1/2}$ (see equation 24). Black shows these moments evaluated using the inferred quasar continuum that results from our joint statistical inference of the PCA coefficients and the astrophysical parameters. Green shows the moments arising from the intrinsic imperfections of the DR, i.e. for the case where we fit our PCA basis (see Fig. 5) to the same set of spectra over the entire spectral range with no noise or IGM absorption. For $\lambda \gtrsim 1220$ Å, the inferred quasar continua (black) achieve the intrinsic limiting precision of the DR itself (green, i.e. PCA fits without noise or IGM absorption), whereas at bluer wavelengths, degeneracy with and censorship by IGM absorption is seen to increase the errors

Figure 15. Efficacy of quasar continuum reconstruction algorithms as quantified by the moments of the relative reconstruction error δ. The lower panels show the mean $\langle\delta\rangle$ (dashed) and standard deviation $\sigma_\delta \equiv \Delta_{\lambda\lambda}^{1/2}$ (solid) of the relative reconstruction error evaluated from $N = 100$ mock spectra on which we performed statistical inference (see Fig. 10 and Section 5.3). Black shows these moments evaluated using the inferred quasar continuum (see equation 32) that results from our joint statistical inference of the PCA coefficients and the astrophysical parameters. Green shows the moments arising from the intrinsic imperfections of the DR alone, i.e. for the case where we fit the PCA basis to the same set of spectra over the entire spectral range with no noise or IGM absorption (see equation 33). For $\lambda > 1220$ Å, the inferred quasar continua (black) achieve the intrinsic limiting precision of the DR itself (green, i.e. PCA fits without noise or IGM absorption), whereas at bluer wavelengths degeneracy with and censorship by IGM absorption is seen to increase the errors as quantified by σ_δ. For comparison, we show the moments of the relative reconstruction error of the red-side continuum fit (red, right) and the blue-side prediction (blue, left) obtained by applying the Davies et al. (2018b) red–blue prediction algorithm to the same set of $N = 100$ mock spectra. Over the wavelength range 1216 Å $< \lambda <$ 1240 Å most critical for measuring IGM damping wings, our inference significantly outperforms the Davies et al. (2018b) algorithm, yielding σ_δ a factor of 1.7–2.5 lower, with an average reduction of σ_δ over this range of a factor of 2.1. For reference, the upper panel shows the mean quasar spectrum $\langle s \rangle$ constructed from our test set with prominent emission lines labelled.

as quantified by σ_δ. For comparison, we show the moments of the relative reconstruction error of the red-side continuum fit (red, right) and the blue-side prediction (blue, left) obtained by applying the Davies et al. (2018b) red–blue prediction algorithm to the same set of $N = 100$ mock spectra. Over the wavelength range 1216 Å $< \lambda <$ 1240 Å most critical for measuring IGM damping wings, our inference significantly outperforms the Davies et al. (2018b) algorithm, yielding σ_δ a factor of 1.7–2.5 lower, with an average reduction of a factor of 2.1 taken over this wavelength range.

To appreciate the significance of this reduction, consider that for a model with $\langle x_{HI} \rangle = 1.0$ and $\log_{10}(t_Q/\text{yr}) = 6$, the mean IGM transmission varies from $\langle t \rangle = 0.49 - 0.91$ over this same range 1216 Å $< \lambda <$ 1240 Å. Heuristically, the signal-to-noise ratio of IGM damping wing absorption for a single spectral pixel is S/N $\sim$ $(1 - \langle t \rangle)/\sigma_\delta$. Averaging this quantity over the range 1216 Å $< \lambda <$ 1240 Å, we find that our inference yields an average statistical significance of 2.9 to be compared to 1.4 for the Davies et al. (2018b) red–blue prediction algorithm. This heuristic S/N likely underestimates the actual improvement, since one optimally combines all the spectral pixels near Lyα when performing a fit, although correlations of the transmission field $\langle t \rangle$ and the covariance of the relative reconstruction error δ make it difficult to quantify the improvement more rigorously. But it is abundantly clear that our approach of jointly fitting for the astrophysical parameters governing the IGM absorption and the latent variables that describe the continuum yields far more accurate continuum reconstructions than red–blue prediction.

6 SUMMARY AND CONCLUSIONS

In this paper, we introduced a new approach for analysing the IGM damping wings that are imprinted on the proximity zones of EoR quasars. Whereas past work has typically forgone the additional constraining power afforded by the blue-side continuum (1216 Å $\lesssim \lambda \lesssim$ 1280 Å) and opted not to model the large correlated IGM transmission fluctuations in the proximity zone ($\lambda \lesssim$ 1216 Å), we derived a single Bayesian likelihood for the entire spectrum allowing us to fit all of the spectral pixels and thus jointly model the fluctuating transmission in the proximity zone, the smooth IGM damping wing signature, and the underlying quasar continuum simultaneously. The latter constitutes a nuisance stochastic process from the standpoint of constraining the average IGM neutral fraction, $\langle x_{HI} \rangle$, and quasar lifetime, t_Q, that govern the IGM transmission. A key aspect of our approach is the use of DR to describe the quasar continuum with a small number of latent variables and then designate the imperfections of this model, which we refer to as relative reconstruction errors, as a source of irreducible correlated noise. Using a large sample of 15 559 SDSS/BOSS quasars at $z \gtrsim 2.15$ we trained and quantified the performance of six distinct DR methods, including machine-learning techniques like GPLVMs and VAEs, and find that a six parameter PCA model (five PCA coefficients ξ plus a normalization s_{norm}) performs best (see also Kist et al. 2025), with complex machine-learning methods providing no advantage. Fitting this PCA model to a subset of 778 spectra which were unseen by the training process

provides an empirical calibration of the relative reconstruction errors, which are an important ingredient of the likelihood we derived. Following our approach, all sources of error – the stochasticity induced by the reionizing IGMs ionization topology, the unknown quasar lifetime t_Q and location of the corresponding ionization front, continuum reconstruction errors, and spectral noise – are accounted for by a generative probabilistic model, which enables us to marginalize out nuisance parameters in a principled manner.

The only drawback of the Gaussian likelihood that we derive in this paper is that it is approximate, because the true likelihood is analytically intractable. We used HMC to conduct statistical inference on an ensemble of 100 realistic mock EoR quasar spectra to determine the coverage of our inference with this approximate likelihood, which quantifies the validity of the credibility contours that we obtain for $\langle x_{HI}\rangle$ and $\log_{10}(t_Q/\mathrm{yr})$ from this new method. We find that our inference is overconfident, which is to say that the 68 per cent credibility contour contains the true astrophysical parameters ($\langle x_{HI}\rangle$ and $\log_{10}(t_Q/\mathrm{yr})$) just 48 per cent of the time, and the 95 per cent credibility contour contains the true parameters just 79 per cent of the time (see Fig. 11). We show convincingly that this overconfidence results from the non-Gaussianity of the IGM transmission field, t, at proximity zone pixels λ where the mean transmission takes on low values, $\langle t_\lambda\rangle \sim 0.1 - 0.2$, owing both to the fact that, physically, the transmission must be positive, and because of an intrinsic strong tail towards higher transmission values (see Fig. 14). Although the HMC posterior samples from our approximate likelihood yield biased inference, we introduced a procedure whereby the HMC samples can be assigned weights, such that the reweighted samples produce reliable inference which passes a coverage test by construction. This reweighting procedure, which amounts to a small dilation of the credibility contours of the original inference, finally yields a state-of-the-art Bayesian inference pipeline that uses all of the spectral pixels to reliably measure the cosmic reionization history and quasar lifetime from quasar spectra.

The accuracy of the quasar continuum reconstructions afforded by this new method are unprecedented. For $\lambda > 1220$ Å, we find that our inferred quasar continua achieve the intrinsic limiting precision of the DR model itself, in other words, they are as good as fits to 'perfect' spectra without noise or IGM absorption. At bluer wavelengths, as expected, degeneracy with and censorship by IGM absorption degrades our ability to reconstruct the underlying quasar continuum. But we nevertheless achieve far more accurate reconstructions than the red–blue prediction algorithms that have been adopted in previous IGM damping wing measurements. Over the wavelength range 1216 Å $< \lambda < 1240$ Å most critical for such measurements, our continuum reconstructions have a factor of 1.7–2.5 smaller error than red–blue prediction, which increases the statistical significance of a putative IGM damping wing per spectral pixel to ~ 2.9 compared to ~ 1.4 for red–blue prediction.

In our companion paper (Kist et al. 2025), we quantify the precision with which IGM damping wings analysed with this new inference approach can measure the astrophysical parameters, $\langle x_{HI}\rangle$ and $\log_{10}(t_Q/\mathrm{yr})$, and the dependence of this precision on the location in parameter space, the dimensionality of the DR latent variable model, as well as on the spectral resolution, S/N, and spectral coverage of the quasar spectra that are analysed. After performing a battery of tests on 1000 mocks, Kist et al. (2025) find that the precision is highest when running this new pipeline with a six-parameter DR model (five PCA coefficients $\boldsymbol{\xi}$ plus a normalization s_{norm}) on S/N ~ 10 spectra, rebinned to a ~ 500 km s^{-1} velocity pixel scale, and extending at least out to the C IV $\lambda 1549$ Å emission line. With this configuration, Kist et al. (2025) find that a single EoR quasar spectrum constrains

the IGM neutral fraction, $\langle x_{HI}\rangle$, to $28.0^{+8.2}_{-8.8}$ per cent and the quasar lifetime, $\log_{10}(t_Q/\mathrm{yr})$, to $0.80^{+0.22}_{-0.55}$ dex, where the error bars indicate the 16 and 84 percentile ranges, and where the constraints improve on both parameters for spectra with a stronger IGM damping wing signature.

Higher precision constraints on $\langle x_{HI}(z)\rangle$ can of course be achieved by averaging over statistical samples of EoR quasars samples. An ambitious program to obtain sensitive *JWST* spectra of the sample of hundreds of EoR quasars that will be delivered by the ESA/*Euclid* satellite would revolutionize the study of IGM damping wings towards quasars and constrain the cosmic reionization history to unprecedented precision. Averaging over ~ 30 quasars in a redshift bin would deliver a precision of ~ 5 per cent on $\langle x_{HI}\rangle$ at that redshift, which, when performed as a function of redshift across the EoR would measure $\langle x_{HI}(z)\rangle$ far more precisely than the CMB. Furthermore, such an analysis would also yield, as a byproduct, the distribution of quasar lifetimes (see e.g. Khrykin et al. 2021) providing novel constraints on the buildup of supermassive black holes (SMBHs) in the young Universe.

Given that the formal precision achievable on $\langle x_{HI}(z)\rangle$ is so high, a natural question arises: will modelling uncertainties due to poorly understood galaxy formation physics eventually limit the precision with which we can constrain reionization? How sensitive are our IGM transmission models (see Section 4) to galaxy formation, which regulates both the ionizing photon sources (via Lyman continuum escape) and sinks (via Lyman limit systems) that determine the reionization topology and the size of the ionized bubble powered by the quasar itself? While galaxy formation physics ultimately determines the reionization topology, we emphasize that precision constraints on $\langle x_{HI}(z)\rangle$ do not require that this topology can be predicted from first principles. Specifically, we adopted a parametrized seminumerical 21CMFAST model (Mesinger et al. 2011; Davies & Furlanetto 2022) in which the source and sink prescriptions are governed by a handful of tunable subgrid parameters. Although we fix these parameters in the present study, yielding a fiducial reionization topology as a function of a single parameter, $\langle x_{HI}(z)\rangle$, an important direction for future work would be to vary and marginalize over the full suite of subgrid source/sink parameters to assess their impact on the inferred $\langle x_{HI}(z)\rangle$ constraints. Furthermore, as discussed below, it is likely that IGM damping wing measurements have the potential to constrain the reionization topology (or the subgrid parameters that govern it) as well (Kist et al., in preparation).

Along similar lines, while our 1D RT of the quasar's radiation currently neglects dense absorbers in the quasar environment (Lyman limit systems and DLAs), which are not captured by the Nyx hydrodynamical simulations, such absorbers can be directly identified in high-S/N quasar spectra via associated metal-line absorption systems (e.g. Davies et al. 2023) and excluded from analysis – analogous to how 'gold samples' are selected in supernova cosmology. Moreover, if optically thick absorbers introduce additional opacity in quasar proximity zones, this can be empirically tested using existing high-resolution spectra of $z \sim 5 - 6$ quasars (for which the global IGM is expected to be highly ionized), for example by comparing the flux PDF of their proximity zones to our models. Any residual disagreement could be addressed by introducing a simple subgrid opacity parameter to govern Lyman limit systems in the 1D RT modelling (e.g. Khrykin et al. 2016; Davies et al. 2016). Thus to summarize, while the impact of uncertain galaxy formation physics on the reionization topology and the presence of dense absorbers in the quasar environment are important considerations that may require adding additional nuisance parameters, they do not constitute a fundamental limitation of our approach. Precision constraints on

reionization do not rest upon a full *ab initio* solution to galaxy formation. Rather, the impact of galaxy formation physics can be treated via a small set of empirically calibratable and marginalizable modelling uncertainties. This situation is closely analogous to precision weak lensing analyses, where the impact of baryonic physics on the matter power spectrum is accounted for using parametric transfer functions derived from hydrodynamical simulations, which are then marginalized over to recover unbiased cosmological constraints (e.g. Heymans et al. 2021; Abbott et al. 2022; Schaller et al. 2024). We therefore argue that IGM damping wings towards quasars provide a viable and powerful method for precision constraints on reionization and SMBH growth, but modelling and marginalizing over uncertainties due to galaxy formation constitutes an important direction for future work.

In addition to addressing these modelling uncertainties, there are several other promising avenues for improving upon the results presented here. First and foremost, our paper argued that an optimal analysis algorithm must construct a fully generative probabilistic model for the entire spectrum, however our likelihood is not optimal because it is only an approximation to the true intractable likelihood. As a result, we had to dilate our credibility contours which degraded the precision of our parameter constraints. Hence, an obvious priority for the future is to attack this problem in the simulation-based inference framework (see e.g. Cranmer, Brehmer & Louppe 2020, for a review) and use machine learning to obtain an expression for the intractable likelihood that we here approximated as a Gaussian (e.g. Chen, Speagle & Rogers 2023). This would surely result in higher precision parameter constraints, both because coverage correction would not be needed and because only the true likelihood can achieve the true optimal precision. Finally, our companion paper (Kist et al. 2025) finds that roughly half of the error budget on $\langle x_{\rm HI} \rangle$ originates from variations of IGM damping wing strength arising from the stochastic distribution of the line-of-sight neutral column density resulting from the topology of reionization itself (Davies et al. 2018c). Several studies have recently noted that IGM damping wing transmission profiles are actually well described by a single parameter (Chen et al. 2024) which is effectively this line-of-sight neutral column density $N_{\rm HI}$ (Keating et al. 2024b; Kist et al., in preparation). It follows that models of IGM damping wings can be parametrized in two distinct ways – either one elects to measure the volume-averaged neutral fraction $\langle x_{\rm HI} \rangle$, which in turn governs the stochastic distribution of the line-of-sight $N_{\rm HI}$ (via an assumed reionization topology), or one can measure an $N_{\rm HI}$ for each quasar individually, and use ensembles of quasars to map out the distribution of $N_{\rm HI}$ empirically as a function of redshift. Whereas this paper adopted the former formulation, an upcoming study explores the latter using the inference machinery that we developed here (Kist et al., in preparation). The great advantage of this latter approach is that it opens up the exciting possibility of using ensembles of quasars to actually determine the distribution of a ~ 100 cMpc 1D moment through the Universe's $n_{\rm HI}$ field, which would not only measure the Universe's reionization history, $\langle x_{\rm HI}(z) \rangle$, but also possibly constrain its topology. These are the primary objectives of cosmological studies of reionization in general and 21-cm experiments in particular, and the methodology that we have presented here paves the way for achieving them with EoR quasar spectra.

DATA AVAILABILITY

The derived data generated in this research will be shared on reasonable requests to the corresponding author.

REFERENCES

Abbott T. M. C. et al., 2022, Phys. Rev. D, 105, 023520
Almgren A. S., Bell J. B., Lijewski M. J., Lukić Z., Van Andel E., 2013, ApJ, 765, 39
Astropy Collaboration, 2013, A&A, 558, A33
Astropy Collaboration, 2018, AJ, 156, 123
Astropy Collaboration, 2022, ApJ, 935, 167
Bañados E. et al., 2018, Nature, 553, 473
Baydin A. G., Pearlmutter B. A., Radul A. A., Siskind J. M., 2018, J. Mach. Learn. Res., 18, 1
Becker G. D., Hewett P. C., Worseck G., Prochaska J. X., 2013, MNRAS, 430, 2067
Becker G. D. et al., 2019, ApJ, 883, 163
Becker G. D., D'Aloisio A., Christenson H. M., Zhu Y., Worseck G., Bolton J. S., 2021, MNRAS, 508, 1853
Becker G. D., Bolton J. S., Zhu Y., Hashemi S., 2024, MNRAS, 533, 1525
Betancourt M., 2017, preprint (arXiv:1701.02434)
Bingham E. et al., 2018, preprint (arXiv:1810.09538)
Bingham E. et al., 2019, J. Mach. Learn. Res., 20, 28:1
Bolton J. S., Haehnelt M. G., 2007a, MNRAS, 374, 493
Bolton J. S., Haehnelt M. G., 2007b, MNRAS, 381, L35
Bolton J. S., Haehnelt M. G., 2013, MNRAS, 429, 1695
Bolton J. S., Haehnelt M. G., Warren S. J., Hewett P. C., Mortlock D. J., Venemans B. P., McMahon R. G., Simpson C., 2011, MNRAS, 416, L70
Boroson T. A., Green R. F., 1992, ApJS, 80, 109
Bosman S. E. I., Kakiichi K., Meyer R. A., Gronke M., Laporte N., Ellis R. S., 2020, ApJ, 896, 49
Bosman S. E. I., Ďurovčíková D., Davies F. B., Eilers A.-C., 2021, MNRAS, 503, 2077
Bradbury J. et al., 2018, JAX: Composable transformations of Python + NumPy programs, Github, available at: http://github.com/google/jax
Byler N., Dalcanton J. J., Conroy C., Johnson B. D., 2017, ApJ, 840, 44
Carswell R. F., Whelan J. A. J., Smith M. G., Boksenberg A., Tytler D., 1982, MNRAS, 198, 91
Cen R., Haiman Z., 2000, ApJ, 542, L75
Chen H., 2024, MNRAS, 528, L33
Chen H., Gnedin N. Y., 2021, ApJ, 911, 60

Chen H. et al., 2022, ApJ, 931, 29

Chen H., Speagle J., Rogers K. K., 2023, preprint (arXiv:2311.16238)

Chen Z., Stark D. P., Mason C., Topping M. W., Whitler L., Tang M., Endsley R., Charlot S., 2024, MNRAS, 528, 7052

Cheng C. et al., 2018, ApJ, 868, 26

Christensen L. et al., 2023, A&A, 680, A82

Collette A., 2013, Python and HDF5. O'Reilly

Cranmer K., Brehmer J., Louppe G., 2020, Proc. Natl. Acad. Sci., 117, 30055

D'Eugenio F. et al., 2024, A&A, 689, A152

Dall'Aglio A., Wisotzki L., Worseck G., 2008, A&A, 491, 465

Davies F. B., Furlanetto S. R., 2022, MNRAS, 514, 1302

Davies F. B., Furlanetto S. R., McQuinn M., 2016, MNRAS, 457, 3006

Davies F. B., Hennawi J. F., Eilers A.-C., Lukić Z., 2018a, ApJ, 855, 106

Davies F. B. et al., 2018b, ApJ, 864, 142

Davies F. B. et al., 2018c, ApJ, 864, 143

Davies F. B., Hennawi J. F., Eilers A.-C., 2019, ApJ, 884, L19

Davies F. B., Hennawi J. F., Eilers A.-C., 2020, MNRAS, 493, 1330

Davies F. B., Bañados E., Hennawi J. F., Bosman S. E. I., 2023, preprint (arXiv:2312.06747)

Davies F. B. et al., 2024, ApJ, 965, 134

Dawson K. S. et al., 2013, AJ, 145, 10

Dawson K. S. et al., 2016, AJ, 151, 44

Delchambre L., 2015, MNRAS, 446, 3545

DESI Collaboration, 2025, J. Cosmol. Astropart. Phys., 2025, 012

Di Valentino E., Melchiorri A., 2022, ApJ, 931, L18

Dijkstra M., Mesinger A., Wyithe J. S. B., 2011, MNRAS, 414, 2139

Doré O., Hennawi J. F., Spergel D. N., 2004, ApJ, 606, 46

Duane S., Kennedy A. D., Pendleton B. J., Roweth D., 1987, Phys. Lett. B, 195, 216

Ďurovčíková D., Katz H., Bosman S. E. I., Davies F. B., Devriendt J., Slyz A., 2020, MNRAS, 493, 4256

Ďurovčíková D. et al., 2024, ApJ, 969, 162

Eilers A.-C., Davies F. B., Hennawi J. F., Prochaska J. X., Lukić Z., Mazzucchelli C., 2017a, ApJ, 840, 24

Eilers A.-C., Hennawi J. F., Lee K.-G., 2017b, ApJ, 844, 136

Eilers A.-C., Hennawi J. F., Davies F. B., 2018, ApJ, 867, 30

Eilers A.-C. et al., 2020, ApJ, 900, 37

Eilers A.-C., Hennawi J. F., Davies F. B., Simcoe R. A., 2021, ApJ, 917, 38

Eilers A.-C., Hogg D. W., Schölkopf B., Foreman-Mackey D., Davies F. B., Schindler J.-T., 2022, ApJ, 938, 17

Euclid Collaboration, 2019, A&A, 631, A85

Fan X., et al., 2006, AJ, 132, 117

Fathivavsari H., 2020, ApJ, 898, 114

Ferraro S., Smith K. M., 2018, Phys. Rev. D, 98, 123519

Foreman-Mackey D., 2016, J. Open Source Softw., 1, 24

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

Gaikwad P. et al., 2023, MNRAS, 525, 4093

Gardner J. R., Pleiss G., Bindel D., Weinberger K. Q., Wilson A. G., 2018, preprint (arXiv:1809.11165)

Garnett R., Ho S., Bird S., Schneider J., 2017, MNRAS, 472, 1850

George E. M. et al., 2015, ApJ, 799, 177

Greig B., Mesinger A., McGreer I. D., Gallerani S., Haiman Z., 2017a, MNRAS, 466, 1814

Greig B., Mesinger A., Haiman Z., Simcoe R. A., 2017b, MNRAS, 466, 4239

Greig B., Mesinger A., Davies F. B., Wang F., Yang J., Hennawi J. F., 2022, MNRAS, 512, 5390

Greig B. et al., 2024a, MNRAS, 530, 3208

Greig B. et al., 2024b, MNRAS, 533, 3312

Harris C. R. et al., 2020, Nature, 585, 357

Heintz K. E. et al., 2024, Science, 384, 890

Heintz K. E. et al., 2025, A&A, 693, A60

Hensman J., Fusi N., Lawrence N. D., 2013, preprint (arXiv:1309.6835)

The HERA Collaboration 2022, ApJ, 925, 221

Heymans C. et al., 2021, A&A, 646, A140

Higgins I., Matthey L., Pal A., Burgess C., Glorot X., Botvinick M., Mohamed S., Lerchner A., 2017, in International Conference on Learning Representations. https://openreview.net/forum?id = Sy2fzU9gl

Hikage C. et al., 2019, PASJ, 71, 43

Hoag A. et al., 2019, ApJ, 878, 12

Hoffman M. D., Gelman A., 2014, J. Mach. Learn. Res., 15, 1593

Hoffman M. D., Blei D. M., Wang C., Paisley J., 2013, J. Mach. Learn. Res., 14, 1303

Horowitz B., Lee K.-G., White M., Krolewski A., Ata M., 2019, ApJ, 887, 61

Hunter J. D., 2007, Comput. Sci. Eng., 9, 90

Jin X. et al., 2023, ApJ, 942, 59

Jung I. et al., 2020, ApJ, 904, 144

Keating L. C., Haehnelt M. G., Cantalupo S., Puchwein E., 2015, MNRAS, 454, 681

Keating L. C., Bolton J. S., Cullen F., Haehnelt M. G., Puchwein E., Kulkarni G., 2024a, MNRAS, 532, 1646

Keating L. C., Puchwein E., Bolton J. S., Haehnelt M. G., Kulkarni G., 2024b, MNRAS, 531, L34

Khrykin I. S., Hennawi J. F., McQuinn M., Worseck G., 2016, ApJ, 824, 133

Khrykin I. S., Hennawi J. F., Worseck G., Davies F. B., 2021, MNRAS, 505, 649

Kingma D. P., Welling M., 2013, preprint (arXiv:1312.6114)

Kirkman D. et al., 2005, MNRAS, 360, 1373

Kist T., Hennawi J. F., Davies F. B., 2025, MNRAS, 538, 2704

Kreisch C. D., Cyr-Racine F.-Y., Doré O., 2020, Phys. Rev. D, 101, 123505

Lalchand V., Ravuri A., Lawrence N. D., 2022, preprint (arXiv:2202.12979)

Lawrence N., 2005, J. Mach. Learn. Res., 6, 1783

Lee K.-G., Suzuki N., Spergel D. N., 2012, AJ, 143, 51

Lee K.-G. et al., 2013, AJ, 145, 69

Lee K.-G. et al., 2015, ApJ, 799, 196

Leschinski K., 2021, Astrophysics Source Code Library, record ascl:2109.007

Liu B., Bordoloi R., 2021, MNRAS, 502, 3510

López S. et al., 2016, A&A, 594, A91

Lukić Z., Stark C. W., Nugent P., White M., Meiksin A. A., Almgren A., 2015, MNRAS, 446, 3697

Lusso E., Worseck G., Hennawi J. F., Prochaska J. X., Vignali C., Stern J., O'Meara J. M., 2015, MNRAS, 449, 4204

Malloy M., Lidz A., 2015, ApJ, 799, 179

Mason C. A., Treu T., Dijkstra M., Mesinger A., Trenti M., Pentericci L., de Barros S., Vanzella E., 2018, ApJ, 856, 2

Mason C. A. et al., 2019, MNRAS, 485, 3947

McGreer I. D., Mesinger A., D'Odorico V., 2015, MNRAS, 447, 499

Mesinger A., Furlanetto S., 2007, ApJ, 669, 663

Mesinger A., Furlanetto S., Cen R., 2011, MNRAS, 411, 955

Mesinger A., Aykutalp A., Vanzella E., Pentericci L., Ferrara A., Dijkstra M., 2015, MNRAS, 446, 566

Mesinger A., Greig B., Sobacchi E., 2016, MNRAS, 459, 2342

Miralda-Escudé J., 1998, ApJ, 501, 15

Morey K. A., Eilers A.-C., Davies F. B., Hennawi J. F., Simcoe R. A., 2021, ApJ, 921, 88

Mortlock D. J. et al., 2011, Nature, 474, 616

Murakami Y. S. et al., 2023, J. Cosmol. Astropart. Phys., 2023, 046

Nikolić I., Mesinger A., Qin Y., Gorce A., 2023, MNRAS, 526, 3170

Oñorbe J., Hennawi J. F., Lukić Z., 2017, ApJ, 837, 106

Pâris I. et al., 2011, A&A, 530, A50

Pâris I. et al., 2017, A&A, 597, A79

Pâris I. et al., 2018, A&A, 613, A51

Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825

Pérez F., Granger B. E., 2007, Comput. Sci. Eng., 9, 21

Phan D., Pradhan N., Jankowiak M., 2019, preprint (arXiv:1912.11554)

Planck Collaboration VI, 2020, A&A, 641, A6

Prochaska J. X., 2017, Astron. Comput., 19, 27

Prochaska J. X., Hennawi J. F., Westfall K. B., Cooke R. J., Wang F., Hsyu T., Davies F. B., Farina E. P., 2020, J. Open Source Softw., 56, 2308

Raiter A., Schaerer D., Fosbury R. A. E., 2010, A&A, 523, A64

Reiman D. M., 2020, preprint (arXiv:2006.00615)

Riess A. G. et al., 2022, ApJ, 934, L7

Sadoun R., Zheng Z., Miralda-Escudé J., 2017, ApJ, 839, 44

Satyavolu S., Kulkarni G., Keating L. C., Haehnelt M. G., 2023, MNRAS, 521, 3108

Schaller M., Schaye J., Kugel R., Broxterman J. C., van Daalen M. P., 2024, preprint (arXiv:2410.17109)

Sellentin E., Starck J.-L., 2019, J. Cosmol. Astropart. Phys., 2019, 021

Shen Y. et al., 2019, ApJ, 873, 35

Sorini D., Oñorbe J., Hennawi J. F., Lukić Z., 2018, ApJ, 859, 125

Spergel D. N. et al., 2003, ApJS, 148, 175

Spina B., Bosman S. E. I., Davies F. B., Gaikwad P., Zhu Y., 2024, A&A, 688, L26

Sun Z., Ting Y.-S., Cai Z., 2023, ApJS, 269, 4

Suzuki N., 2006, ApJS, 163, 110

Suzuki N., Tytler D., Kirkman D., O'Meara J. M., Lubin D., 2005, ApJ, 618, 592

Titsias M., Lawrence N. D., 2010, in Teh Y. W., Titterington M., eds, Proc. Machine Learning Research Vol. 9, Proc. 13th International Conference on Artificial Intelligence and Statistics. PMLR, Sardinia, Italy, p. 844, https://proceedings.mlr.press/v9/titsias10a.html

Titsias M., Lázaro-Gredilla M., 2014, in International Conference on Machine Learning, p. 1971

Umeda H., Ouchi M., Nakajima K., Harikane Y., Ono Y., Xu Y., Isobe Y., Zhang Y., 2024, ApJ, 971, 124

van den Busch J. L. et al., 2022, A&A, 664, A170

Virtanen P. et al., 2020, Nat. Methods, 17, 261

Wang F. et al., 2020, ApJ, 896, 23

Wolfson M., Hennawi J. F., Davies F. B., Oñorbe J., 2023, MNRAS

Yang J. et al., 2020, ApJ, 897, L14

Yip C. W. et al., 2004, AJ, 128, 2603

Young P. J., Sargent W. L. W., Boksenberg A., Carswell R. F., Whelan J. A. J., 1979, ApJ, 229, 891

Zahn O. et al., 2012, ApJ, 756, 65

Zel'Dovich Y. B., 1970, A&A, 500, 13

Zhou Y., Chen H., Di Matteo T., Ni Y., Croft R. A. C., Bird S., 2024, MNRAS, 528, 3730

Zhu Y. et al., 2023, ApJ, 955, 115

Zhu Y. et al., 2024, MNRAS, 533, L49

APPENDIX A: COVERAGE TESTS

In this appendix, we provide details on the coverage test presented in Section 5.2, the reweighting scheme described in Section 5.4, and the application of both to marginal posterior distributions.

A1 An algorithm for performing a coverage test

Below we provide a description of an algorithm for carrying out a coverage test.

(i) Draw N parameter vectors $\mathbf{\Theta}_{\text{true},j}$ from the prior distribution $P(\mathbf{\Theta})$. These are the 'true' parameters that generate the mock data sets used to perform the coverage test.

(ii) Using a forward simulator, generate a set of N mock data sets, $\mathbf{x}_j$, for each of the parameter vectors $\mathbf{\Theta}_{\text{true},j}$.

(iii) Inference is carried out on each data set resulting in a set of N posterior distributions $P(\mathbf{\Theta}|\mathbf{x}_j)$.

(iv) Consider a set of M credibility contour levels $\alpha \in [0, 1]$. For each value α and each mock data set $(\mathbf{\Theta}_{\text{true},j}, \mathbf{x}_j)$, one tests whether the true parameter value $\mathbf{\Theta}_{\text{true},j}$ resides within the volume V_α enclosed by the αth contour, defined by

$$\int_{V_\alpha} P(\boldsymbol{\theta}|\mathbf{x}_j)\mathrm{d}\boldsymbol{\theta} = \alpha. \tag{A1}$$

For each α, the coverage probability $C(\alpha)$ is the fraction of the N mock data sets for which the true value $\boldsymbol{\theta}_{\text{true},j}$ lies within the volume V_α.

The result of the coverage test is the relation $C(\alpha)$ versus α. A perfect inference procedure would yield $C(\alpha) = \alpha$ in the limit

$N \to \infty$. An overconfident inference pipeline will yield a curve $C(\alpha)$ versus α that lies systematically below the line $y = x$, whereas for an underconfident inference procedure the $C(\alpha)$ will lie above $y = x$. This algorithm yields an unbiased estimate $C(\alpha)$ of the underlying coverage probability from a finite set of mock data sets N. Since $C(\alpha)$ counts how often the true parameters fall inside the αth contour, it is the number of successes in a sequence of N independent experiments each asking a yes-or-no question – success occurring with a probability $p = C(\alpha)$ and failure occuring with probability $q = 1 - C(\alpha)$. Thus, by definition our estimate $C(\alpha)$ must follow the binomial distribution $B(N, C(\alpha))$, which can be used to assign error bars to $C(\alpha)$ resulting from the finite number of mock data sets N.

Underlying this coverage test algorithm is a procedure for testing whether a true parameter vector resides within the volume V_α enclosed by a contour corresponding to credibility level α. A contour of the posterior $P(\mathbf{\Theta}|\mathbf{x}_j)$ containing a fraction α of the total probability slices the posterior along an isodensity level P_α, such that the volume V_α in equation (A1) is defined by

$$V_\alpha = \{\mathbf{\Theta} \mid P(\mathbf{\Theta}|\mathbf{x}_j) \geq P_\alpha\}. \tag{A2}$$

A parameter value $\mathbf{\Theta}'$ will lie within the volume enclosed by the αth contour provided that $P(\mathbf{\Theta}'|\mathbf{x}_j) \geq P_\alpha$, so the procedure boils down to estimating the set of isodensity levels P_α corresponding to the set of credibility levels α.

In practice, one typically has a number of samples, n, from the posterior from a run of an MCMC or HMC sampling algorithm. If we rewrite equation (A1) as an integral over the entire parameter space $\mathbf{\Theta}$:

$$\int H[P(\mathbf{\Theta}|\mathbf{x}_j) - P_\alpha]P(\mathbf{\Theta}|\mathbf{x}_j)\mathrm{d}\mathbf{\Theta} = \alpha, \tag{A3}$$

where the Heaviside function, $H(x)$, enforces the condition on the volume from equation (A2), then it can be evaluated via Monte Carlo integration according to

$$\int H[P(\mathbf{\Theta}|\mathbf{x}_j) - P_\alpha]P(\mathbf{\Theta}|\mathbf{x}_j)\mathrm{d}\mathbf{\Theta} = \frac{1}{n}\sum_i^n H[P(\mathbf{\Theta}_i|\mathbf{x}_j) - P_\alpha], \tag{A4}$$

where the sum is over the n MCMC or HMC posterior samples. Hence, we can determine the isodensity level, P_α, corresponding to each credibility level, α, by solving the equation:

$$\begin{aligned}
\frac{1}{n}&\sum_i^n H[P(\mathbf{\Theta}_i|\mathbf{x}_j) - P_\alpha] \\
&= \frac{\text{Number of samples with } P(\mathbf{\Theta}|\mathbf{x}_j) \geq P_\alpha}{n} = \alpha \\
&= \text{CDF}(\geq P_\alpha) = \alpha. \tag{A5}
\end{aligned}$$

The second equality in equation (A5) indicates that credibility contour definition amounts to computing the cumulative probability distribution, $\text{CDF}(\geq P_\alpha)$, of the posterior distribution at the MCMC or HMC samples, $\mathbf{\Theta}_i$. One can then invert the CDF

$$P_\alpha = \text{CDF}^{-1}(\alpha), \tag{A6}$$

to determine the corresponding isodensity levels. This procedure can be employed to test each of the N posteriors in the inference test for all credibility levels. If the true model parameters, $\mathbf{\Theta}_{\text{true},j}$, lie inside the volume of credibility contour α then the condition $P(\mathbf{\Theta}_{\text{true},j}|\mathbf{x}_j) \geq P_\alpha$ will be satisfied. MCMC/HMC samplers typically can return the value of $P(\mathbf{\Theta}_i|\mathbf{x})$ at every sample in the chain, allowing one to

easily estimate CDF($\geq P_\alpha$). Then, one only needs to evaluate the $P(\boldsymbol{\Theta}_{\mathrm{true},j}|\boldsymbol{x}_j)$, which is straightforward since the function $P(\boldsymbol{\Theta}|\boldsymbol{x}_j)$ is a pre-requisite for performing inference.

A2 Reweighting posterior samples to pass a coverage test

Consider a scenario where our current inference pipeline fails a coverage test because of some imperfection in our probabilistic model of the measurement process which it is not straightforward to correct. For concreteness, imagine we adopted an approximate form for the posterior distribution because the true form is not analytically tractable, and that our coverage test indicates these approximate posteriors are overconfident, which is to say that $C(\alpha)$ versus α lies systematically below the line $y = x$. How can we nevertheless perform statistically reliable inference? Sellentin & Starck (2019) advocate that one simply relabel the contours to reflect this unfortunate reality. For example, if the 68th percentile credibility contour, $\alpha = 0.68$, actually only contains the true model 48 per cent of the time, i.e. $C(\alpha) = 0.48 < \alpha$, then label this contour as the 48th percentile rather than the 68th. The real 68th percentile contour containing the true parameters 68 per cent of the time under the inference test might say correspond to the value $\alpha = 0.85 = C^{-1}(0.68)$ contour for the original approximate inference, which in turn maps to a lower isodensity level of the approximate posterior $P(\boldsymbol{\theta}|\boldsymbol{x})$, i.e. $P_{0.85} < P_{0.68}$. In other words, by choosing a lower isodensity threshold P_α we expand the contours to contain the true model the empirically correct fraction of the time.

In general, this remapping of the credibility levels α into true coverage probabilities $C(\alpha)$ can be achieved by assigning a set of weights to the samples from the approximate posterior. The purpose of MCMC or HMC samples from a posterior is to estimate credibility intervals on parameters, perform marginalization integrals, and compute 'moments' of the posterior via Monte Carlo integration. If we can determine the set of weights that corrects the imperfect inference such that it passes an inference test, these weights can then be used in all of the downstream computations that one performs with the samples. To achieve this we generalize equation (A5) to the case of defining contour levels from reweighted samples

$$\sum_i^n w_i H[P(\boldsymbol{\Theta}_i|\boldsymbol{x}_j) - P'_\alpha] = \mathrm{wCDF}(\geq P'_\alpha) = \alpha, \tag{A7}$$

where we have simply absorbed the constant $1/n$ normalization factor into the definition of the weights. Here, wCDF($\geq P'_\alpha$) is the weighted cumulative distribution function of the original posterior evaluated at the samples, $\boldsymbol{\Theta}_i$. Similar to before, the isodensity levels of the reweighted posterior can be determined according

$$P'_\alpha = \mathrm{wCDF}^{-1}(\alpha). \tag{A8}$$

In order for the weighting to correct the inference for e.g. credibility level $\alpha = 0.68$, we need the solution to equation (A8) to yield $P'_{0.68} = P_{0.85}$ where $0.85 = C^{-1}(0.68)$ and $P_{0.85} = \mathrm{CDF}^{-1}(0.85)$ is the solution for the isodensity threshold under the original approximate inference from equation (A5). Furthermore, this remapping must hold for all of the isodensity levels $P'_\alpha = P_{C^{-1}(\alpha)} = \mathrm{CDF}^{-1}(C^{-1}(\alpha))$.

To build intuition, first consider the situation where $C(\alpha) = \alpha$, i.e. where we pass the inference test perfectly. In this limiting case, it is clear that the weighting should be uniform and hence $w_i = 1/n$. Without loss of generality we can sort the samples in the sum in equation (A7) in order of increasing $P(\boldsymbol{\Theta}_i|\boldsymbol{x}_j)$, adopting the convention that $P(\boldsymbol{\Theta}_1|\boldsymbol{x}_j)) \leq P(\boldsymbol{\Theta}_2|\boldsymbol{x}_j)) \leq \cdots \leq P(\boldsymbol{\Theta}_n|\boldsymbol{x}_j)$ and the corresponding weights are similarly ordered such that w_1 is the

weight assigned to the sample with smallest value $P(\boldsymbol{\Theta}_1|\boldsymbol{x}_j)$, etc. Since we need to solve for n weights, w_i, we need the same number of values of α as constraints. If we choose n linearly decreasing values of α spanning the range $\alpha = [1, 1/n]$, then equation (A7) implies the following set of linear equations:

$$\begin{aligned} w_1 + w_2 + \ldots \; + w_{n-1} + w_n &= \quad 1 \\ w_2 \; + \cdots + \; w_{n-1} + w_n &= 1 - 1/n \\ \vdots \qquad \vdots \qquad \vdots \quad \vdots \qquad & \\ w_{n-1} + w_n &= \quad 2/n \\ w_n &= \quad 1/n, \end{aligned} \tag{A9}$$

or equivalently in matrix form

$$\mathbf{U}\mathbf{w} = \boldsymbol{\alpha}, \tag{A10}$$

where $\mathbf{U}$ is an upper triangular matrix with all non-zero elements equal to unity, $\mathbf{w}$ is a vector of weights to be assigned to the samples, and $\boldsymbol{\alpha}$ is the vector of uniformly spaced credibility levels. Since the $\boldsymbol{\alpha}$ constraints that we chose can also be interpreted as the cumulative distribution of the posterior evaluated at the samples, i.e. $\alpha_i = \mathrm{CDF}(\geq P_i)$, equation (A9) is equivalent to

$$\mathrm{wCDF}(\geq P_i) = \mathrm{CDF}(\geq P_i). \tag{A11}$$

It is thus no surprise that the solution to this linear system is obviously $w_i = 1/n$ yielding the uniform weighting we expect.

To generalize to the case $C(\alpha) \neq \alpha$, we solve for the weights that satisfy

$$\mathbf{U}\mathbf{w} = \mathbf{C}(\boldsymbol{\alpha}), \tag{A12}$$

where $\mathbf{C}(\boldsymbol{\alpha})$ is now a vector of coverage values evaluated at the vector of uniformly spaced credibility levels $\boldsymbol{\alpha}$. To see why this works consider for example the kth equation in the linear system in equation (A12) where k is chosen to be the sum over the 85th percentile highest rank weights and $P(\boldsymbol{\Theta}_i|\boldsymbol{x}_j)$ values (recall the weights are sorted in order of increasing $P(\boldsymbol{\Theta}_i|\boldsymbol{x}_j)$)

$$w_k + w_{k+1} + \cdots + w_n = C(0.85) = 0.68. \tag{A13}$$

or equivalently

$$\mathrm{wCDF}(\geq P'_{0.68}) = 0.68. \tag{A14}$$

Because we chose $\boldsymbol{\alpha}$ to be linearly decreasing, the kth ranked sample will correspond to the 85th percentile of the CDF of the original approximate posterior values evaluated at the MCMC/HMC samples, and hence it is guaranteed that $P'_{0.68} = P_{0.85} = \mathrm{CDF}^{-1}(0.85)$. In other words, the 68th percentile contour of the new reweighted posterior distribution will correspond to the 85th percentile contour of the original approximate posterior, as desired. The rest of the equations in the linear system similarly enforce the constraints that $P'_\alpha = P_{C^{-1}(\alpha)} = \mathrm{CDF}^{-1}(C^{-1}(\alpha))$ for the other credibility levels α. Since the determinant of an upper triangular matrix is simply the product of the diagonal elements, $\det \mathbf{U} = 1$, and the linear system in equation (A12) will always yield a unique solution for the vector of weights $\mathbf{w}$ which imposes these constraints.

A3 Coverage tests for marginal distributions

Performing inference requires that one can evaluate $P(\boldsymbol{\Theta}|\boldsymbol{x}_j)$ at any location within the parameter space. Furthermore, MCMC and HMC samplers typically return the value of $P(\boldsymbol{\Theta}|\boldsymbol{x}_j)$ at every sample in the chain, allowing one to easily determine the set of isodensity levels for each posterior and test $P(\boldsymbol{\Theta}_{\mathrm{true},j}|\boldsymbol{x}_j) \geq P_\alpha$. But what about

performing inference tests on marginal distributions? To make the discussion more concrete let us imagine that our parameter vector Θ can be split into a set of θ physical parameters of interest and a set of η nuisance parameters. It is easy to see that one can pass a coverage test for the full posterior $P(\Theta|x_j)$ but nevertheless fail a coverage test for the marginal posterior $P(\theta|x_j)$.[21] However, the marginal physical parameter posterior, $P(\theta|x_j)$, is what we actually care about, whereas we would be more willing to tolerate overconfident (or underconfident) total posteriors and marginal nuisance parameter posteriors.

The same coverage test algorithm described in Appendix A1 applies to the marginal case. Specifically, all of the steps of the procedure are the same with the exception of step (iv), where one must now test whether the true physical parameter values $\theta_{\text{true}, j}$ lie within contours of the marginal posterior $P(\theta|x_j)$ for each data set. There is however an important technical difference. The procedure outlined in equation (A5) for determining isodensity levels and testing whether the true physical parameters lie within volumes enclosed by the credibility contours specified by α presumes that one can evaluate $P(\theta|x_j)$ at every sample in the MCMC/HMC chain as well as at the true parameter location. However, in practice evaluating $P(\theta|x_j)$ would require performing a typically intractable marginalization integral over the nuisance parameters. Instead, this intractable marginalization can be performed via Monte Carlo integration. By applying a density estimation algorithm (i.e. histogram,

kernel density estimation, or Gaussian mixture model) to the MCMC or HMC samples marginalized over the nuisance parameters, an expression for $P(\theta|x_j)$ can then be determined allowing one to test whether the true physical parameter values lie within the contours of the marginal posteriors. While this Monte Carlo integration plus density estimation procedure may sound complex, it is exactly the procedure adopted when making a corner plot of MCMC/HMC samples.

---

[21]Consider a thought experiment where there is a single physical parameter and a million nuisance parameters. Imagine a trivial true posterior independently described by $\mathcal{N}(\theta_i|0, 1)$ for each element of Θ. In conducting inference, suppose we use the correct unit variance for each of the million elements of η, but erroneously adopt $\sigma_\theta = 1/2$ for the Gaussian describing θ. It is obvious that the incorrect form for the θ component will be diluted by contributions from the million nuisance parameters, and thus have a negligible impact on the total posterior. Within reasonable numerical precision the incorrect form of $P(\Theta|x_j)$ with $\sigma_\theta = 1/2$ will pass a coverage test. But since the marginalization integral over the million nuisance parameters here trivially evaluates to unity, the marginal posterior is $P(\theta|x_j) = \mathcal{N}(\theta|0, 1/2)$, and it is manifestly clear that it will fail the marginal coverage test.

This paper has been typeset from a TeX/LaTeX file prepared by the author.