Maximally-informative inference from spectroscopic quasar surveys

Keir K. Rogers; hosts: Justin Alsing, Stephen Feeney

One of the most important challenges for astronomy is to find an efficient, yet maximally-informative way to test physical models with the large amounts of data available from astronomical surveys. This is especially true for surveys of large-scale structure, where the physics is non-linear and non-Gaussian and the datasets are large and non-uniform across the sky. Large-scale structure surveys observe tracers of the underlying dark matter distribution. Quasars are tracers in themselves; their spectra contain multiple series of absorption lines which are tracers also (via the intergalactic gas that causes the absorption). The Lyman-alpha forest, the series of Lyman-alpha absorption lines from neutral hydrogen in the intergalactic medium (IGM), is a sensitive probe of structure on a wide range of scales (and redshifts). Current approaches measure two-point correlations on small scales (up to 4 / Mpc) only in the radial direction (i.e. along the lengths of spectra). This small-scale one-dimensional flux power spectrum is particularly constraining of extended cosmological models with additional components, e.g. massive neutrinos or warm dark matter. Correlations in the transverse direction (i.e. between different spectra) are only included on the very largest scales (hundreds of Mpc) principally in order to detect the baryon acoustic oscillation feature, which provides constraints on the expansion history of the Universe.

However, it is known that there exists additional information in the three-dimensional correlation function up to the smallest scales, which is currently not accessed. It is also known that there exists information in the higher-order correlations in the Lyman-alpha forest, which can, e.g., break degeneracies between parameters describing the thermal history of the IGM. Correlations in and between other absorption line series (e.g. the Lyman-beta and triply-ionised carbon (CIV) forests) and cross-correlation with the background quasar distribution have been measured to a limited extent. This project aims to train artificial neural networks to learn non-linear functionals of input data that maximise the Fisher information: information maximising neural networks (IMNNs; Charnock et al., 2018). By inputting ensembles of raw quasar spectra, the IMNN will learn the set of summary statistics that maximally informs about the parameters of a given model. In this way, it will automatically learn the best combination of correlations between the different tracers present in quasar spectra that will most constrain cosmological and astrophysical model parameters.

This process is a data compression, taking the large input dataset of a spectroscopic quasar survey and returning a small number of summary statistics. This permits the use of methods for likelihood-free inference, where comparison between data and model takes place in data-space and likelihood function approximations are avoided. This is particularly useful for large-scale structure surveys, where it is much easier to model complex physical processes, instrumental effects and selection biases in a forward simulation, than to construct a complicated likelihood function and solve the inverse problem. In order to move beyond only measuring correlations within individual spectra and to include the small-scale transverse correlations between quasar sightlines, the mitigation of selection biases will become far more important. This project aims to use the likelihood-free methods of DELFI (density-estimation likelihood-free inference; Alsing et al., 2018), where neural networks learn a Gaussian mixture model of the joint probability distribution of data and parameters, using only forward simulations of quasar spectra as draws from the prior distribution. A complication in the case of quasar spectra is that accurate forward modelling requires the computation of a cosmological hydrodynamical simulation. These are sufficiently computationally expensive that a realistic number of forward simulations available for a given analysis is on the order of 50 - 100. The number of samples required for accurate density estimation is on the order of 10^4. This necessitates the incorporation of methods for interpolating between simulations using Gaussian process emulation and the Bayesian optimisation of forward simulation training sets. The first steps of this project were all carried out during my visit. The plan of the project was agreed, for which the first part is to get DELFI working with the emulator code and a test set of hydrodynamical simulations. Initially, we will use the score-function data compression of Alsing & Wandelt (2017), for which code has been written. Future work will incorporate the IMNNs and a more realistic suite of simulations. I also had very productive discussion with Ben Wandelt, where we outlined how to adapt the existing IMNN architectures to analyse 3D mock sets of quasar spectra and how to deal with the limited number of realisations available for Fisher matrix estimation. This is all in spite of the potential disasters of an exploding asbestos pipe at the Flatiron (necessitating a temporary move to nearby NYU) and a mystery illness to my host, Justin, which delayed his arrival. Nonetheless, the visit would appear to be a success, with the outline of at least one journal publication agreed and significant progress in the necessary work.

This work was supported by collaborative visits funded by the Cosmology and Astroparticle Student and Postdoc Exchange Network (CASPEN).