matryoshka: A Suite of Neural-Network-Based Emulators

Jamie Donald-McCann | jamie.donald-mccann@port.ac.uk

Institute of Cosmology and Gravitation

1. Galaxy Clustering

- Studying to what extent galaxies are *clustered* together throughout the Universe can reveal a great deal of information about the astrophysical processes that define their evolution and the cosmological context they exist in.
- The observed galaxy distribution is often summarised by statistics like the *two-point correlation function*, or its Fourier space equivalent the *power spectrum*. Figure 1 shows an example of the power spectrum *multipoles*; these multipoles are a result of a convolution of the 2D power spectrum measured from BOSS galaxies and Legendre polynomials $\mathcal{L}_l$.
- **Comparing theoretical predictions for summary statistics like the correlation function and the power spectrum to those from observations allows us to put constraints on a cosmological model to describe the observed Universe.**

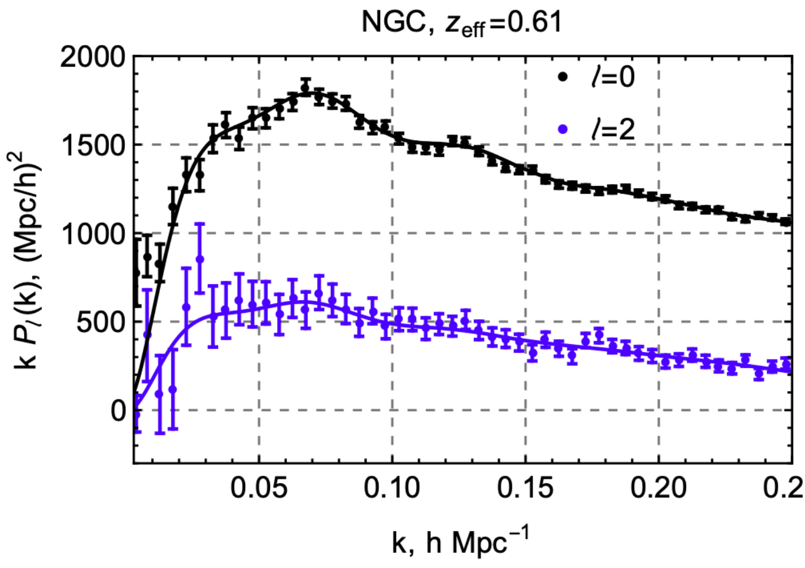


Fig. 1: Power spectrum multipoles measured from northern galactic cap (NGC) BOSS data. The points with errorbars show the measurement, whilst the solid lines show a best-fit model. Plot from [4]

2. Parameter Inference

- The goal when carrying out parameter inference is to find the *posterior distribution* $P(\mathbf{\Phi}|\mathbf{x})$. With $\mathbf{\Phi}$ being the parameters of our theoretical model, and $\mathbf{x}$ being our observed data.
- Generally $P(\mathbf{\Phi}|\mathbf{x})$ is calculated via sampling techniques such as *Markov Chain Monte Carlo* (MCMC).
- The number of theoretical model parameters for astrophysical and cosmological problems can be large, and many of these parameters can be highly correlated. In such cases many samples, and thus many model evaluations, are needed for MCMC chains to reach convergence.
- **For all but the most efficient theoretical models parameter inference is expensive!**

[1] Guido D'Amico et al. In: *JCAP* (May 2020). DOI: 10.1088/1475-7516/2020/05/005.

[2] Jamie Donald-McCann, Kazuya Koyama, and Florian Beutler. In: *Preprint* (Feb. 2022). DOI: 10.48550/arXiv.2202.07557.

[3] Katrin Heitmann et al. In: *ApJ* (Nov. 1, 2009). DOI: 10.1088/0004-637X/705/1/156.

[4] Mikhail M. Ivanov, Marko Simonović, and Matias Zaldarriaga. In: *JCAP* (May 2020). DOI: 10.1088/1475-7516/2020/05/042.

[5] Minas Karamanis, Florian Beutler, and John A. Peacock. In: *MNRAS* 508 (Oct. 2021). DOI: 10.1093/mnras/stab2867.

[6] TensorFlow. Version v2.6.0-rc1. 2021. DOI: 10.5281/ZENODO.4724125.

3. EFTofLSS

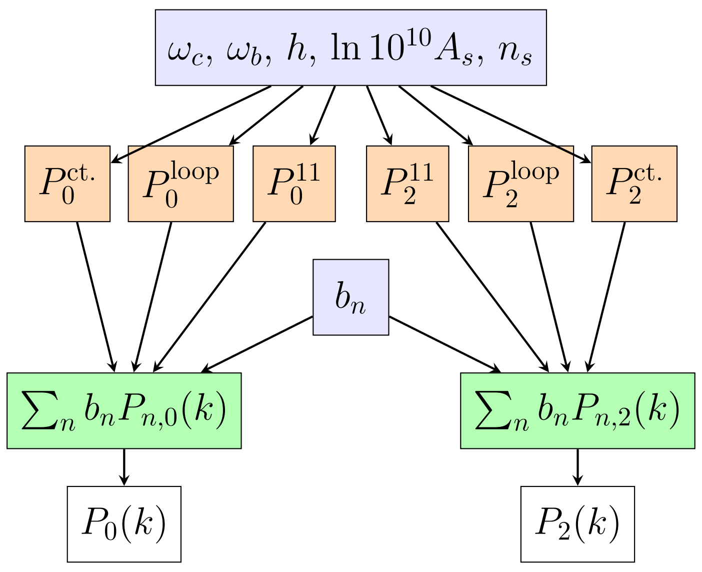
- ***The Effective Field Theory of Large Scale Structure* (EFTofLSS) is a model for galaxy clustering based on *perturbation theory*** that has had significant development over recent years, and has been used to constrain cosmological parameters from BOSS and eBOSS data (see for example [1]).
- **EFTofLSS predictions can be represented as a combination of a series of *bias parameters* b_n that do not depend on cosmology, with a series of components $P_{n,l}$ that do depend on cosmology.** Such that the power spectrum multipoles are given by

$$P_l(k) = \sum_n b_n P_{n,l}(k)\ .$$

- `PyBird` is a public implementation of the EFTofLSS model. Predictions for the monopole and quadrupole of the galaxy power spectrum can be made in $\mathcal{O}(1s)$. This is fast enough to make inference via MCMC possible, but slow enough to mean that significant resource would be required (likely time on a HPC cluster).

4. Cosmic Emulation

- *Emulation* is a technique that was first proposed to cope with the large computational cost of cosmological simulations, and indirectly use their outputs to do cosmological inference [3].
- **An emulator can be thought of as a sophisticated interpolation scheme that aims to rapidly reproduce the outputs of a computationally expensive model. The emulator then acts as a surrogate model which is evaluated when carrying out inference.**
- Simple *neural networks* (*multilayer perceptrons* with small numbers of layers) make good emulators as they can produce vector outputs, they're fast to train and evaluate, and they are capable of approximating smooth functions.
- **With `TensorFlow` [6] we develop a suite of six neural-network-based emulators to approximate the bias-independent $P_{n,l}$ components of the EFTofLSS model [2].** Emulating these $P_{n,l}$ rather than the multipoles directly is advantageous for two main reasons; the dimensionality of each emulator is reduced, and there is no fixed prior on the bias parameters. Figure 2 visualises how the six component emulators are used together to predict the multipoles.



$\omega_c,\ \omega_b,\ h,\ \ln 10^{10}A_s,\ n_s$

$P_0^{\rm ct.}$ $P_0^{\rm loop}$ P_0^{11} P_2^{11} $P_2^{\rm loop}$ $P_2^{\rm ct.}$

b_n

$\sum_n b_n P_{n,0}(k)$ $\sum_n b_n P_{n,2}(k)$

$P_0(k)$ $P_2(k)$

Fig. 2: Flowchart visualising how each of the component emulators is used together when making predictions for the monopole $P_0(k)$ and quadrupole $P_2(k)$ of the galaxy power spectrum. Pale blue boxes represent inputs, orange boxes component emulators, green boxes analytic operations, and white boxes outputs. Each emulator predicts a group of $P_{n,l}$ components that are then combined with the bias parameters b_n analytically.

5. Accelerated Inference

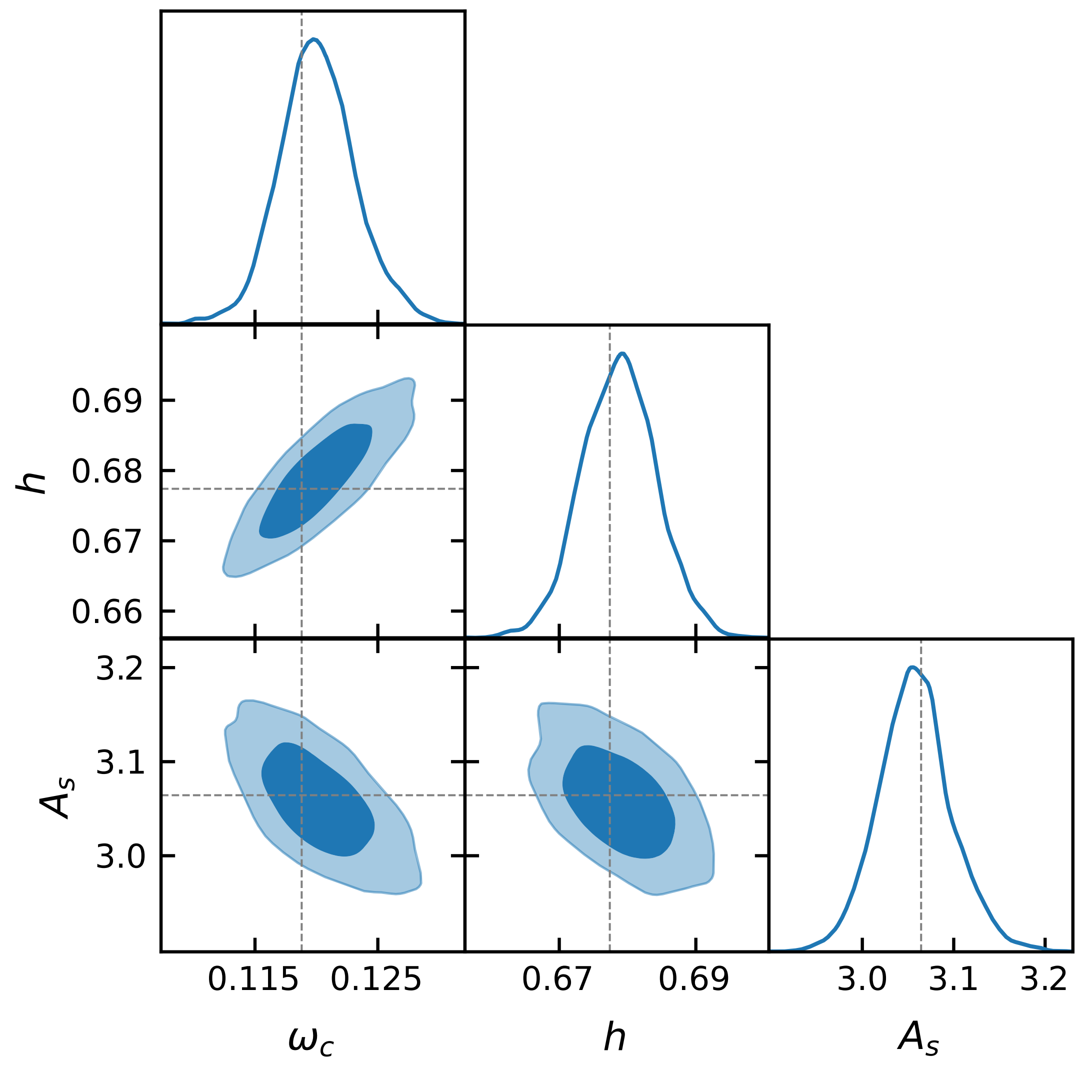


Fig. 3: Corner plot showing posterior probability distributions from mock analysis using suite of EFTofLSS componenet emulators. Grey dashed lines show the locations of the true cosmological parameters of the mocks. The contour levels represent 1σ and 2σ

- **We can test the accuracy of the EFTofLSS component emulators by running mock cosmological parameter inference with data coming from high resolution cosmological simulations.**
- The mock multipoles are measured from a *halo occupation distribution* (HOD) mock; produced with a DESI ELG like HOD and simulation box with 3 Gpc h^{-1} sides. The covariance matrix is calculated from a set of approximate mocks.
- The MCMC is run with an ensemble slice sampling code: `zeus` [5]. Thanks to `TensorFlow` optimisation the emulators are efficient when making batch predictions. To fully exploit this we run the MCMC with a large number of walkers (560 for 14 free parameters) after burn-in is complete.
- Figure 3 shows the posterior distributions of some of the relevant cosmological parameters. **The truth (grey dashed lines) is well recovered within 1σ; indicating any prediction errors are not large enough to induce statistically significant biases in the constrained cosmology.**
- **The MCMC required $\mathcal{O}(10^6)$ model evaluations, and took $\sim$ 30 minutes on a laptop. To generate the same number of samples with `PyBird` would take $\sim$ 2 weeks on the same laptop.**

LaTeX TikZposter