

Error Computation for Emulation

April 6, 2020

1 Introduction and Notation

This document outlines the error measurements used in our emulation procedure.

The test set contains mock catalogs with $C = 7$ different cosmologies c . There are $B = 5$ boxes (or realizations) b for each cosmology. These are each populated with $H = 100$ different halo occupation distribution (HOD) models h , the same 100 for each cosmology. (There are also 10 random seeds used per HOD, but for now we only use one. We'll worry about this later.)

We are concerned with a clustering statistic y . We use y_{meas} to mean the statistic computed on a mock or data. We use y_{pred} to mean the emulator prediction for that statistic. We take y to have P dimensions (bins in our case), indexed by p (or p').

The statistic is computed on all $7 \times 5 \times 100 = 3500$ mocks. The statistic is averaged over the 5 boxes, so our final test set contains $N = 700$ statistics. We will denote these with index n .

2 Test set error

The error on the test set is computed as follows. For each cosmology, compute the mean statistic over the boxes. Compute the deviation from this mean for each of the boxes, defined as the fractional error between the box statistic and the mean statistic for that cosmology. Do this for every HOD. The test set error is the standard deviation of all of these deviations from the mean.

$$\bar{y}_{c,h} = \frac{1}{B} \sum_b^B y_{b,c,h} \quad (1)$$

$$d_{c,h} = \frac{y_{b,c,h} - \bar{y}_{c,h}}{\bar{y}_{c,h}} \quad (2)$$

$$\bar{d} = \frac{1}{CH} \sum_c^C \sum_h^H d_{c,h} \quad (3)$$

$$\sigma_{test} = \sqrt{\frac{1}{CH} \sum_c^C \sum_h^H (d_{c,h} - \bar{d})^2} \quad (4)$$

This test set error σ_{test} is used as an input to the Gaussian Process to represent the error on the *training set*. We do this because have a better handle on this error due to the multiple realizations of the test set. We note that the procedure outlined here estimates the sample variance contribution to the error. The random seeds we have not yet included estimate the shot noise contribution; we will update with this later.

3 Emulator performance and error

The overall emulator performance σ_{perf} is defined as

$$\sigma_{perf}^2 = \sigma_{emu}^2 + \sigma_{test}^2 \quad (5)$$

We obtain this with the RMS error of the test set emulation:

$$\sigma_{perf} = \sum_t^N \sqrt{\frac{(y_{n,pred}^2 - y_{n,meas}^2)}{N}} \quad (6)$$

We care about the emulator error σ_{emu} separately, so we use

$$\sigma_{emu} = \sqrt{\sigma_{perf}^2 - \sigma_{test}^2} \quad (7)$$

4 Covariance and likelihood

The covariance between the bins of the statistic is obtained with the Minerva mock catalogs. We compute the statistic on $M = 100$ Minerva mocks, which we index by m . The covariance matrix C is, in both vector notation and element notation for clarity,

$$C = \frac{1}{N-1} \sum_m^M (y_m - \bar{y})(y_m - \bar{y})^T \quad (8)$$

$$C_{pp'} = \frac{1}{N-1} \sum_m^M (y_{m,p} - \bar{y}_p)(y_{m,p'} - \bar{y}_{p'}) \quad (9)$$

The correlation matrix R is the normalized covariance matrix, with elements

$$R_{pp'} = \frac{C_{pp'}}{\sqrt{C_{pp}C_{p'p'}}} \quad (10)$$

We combine these errors to obtain a final covariance matrix \mathcal{C} for our likelihood, where y_{meas} is the measured statistic on our data:

$$\sigma_{comb,p} = \sqrt{(y_{meas,p}\sigma_{test,p})^2 + \sigma_{emu,p}^2} \quad (11)$$

$$\sigma_{comb,p'} = \sqrt{(y_{meas,p'}\sigma_{test,p})^2 + \sigma_{emu,p'}^2} \quad (12)$$

$$\mathcal{C}_{pp'} = R_{pp'}\sigma_{comb,p}\sigma_{comb,p'} \quad (13)$$

The log-likelihood \mathcal{L} is then defined as

$$\mathcal{L} = -\frac{1}{2}(y_{pred} - y_{meas})\mathcal{C}^{-1}(y_{pred} - y_{meas})^T \quad (14)$$