Program for Statistics and computational tools for Cosmology (80 hours)

Mariana Vargas Magaña (IF-UNAM) and Sebastien Fromenteau (ICF-UNAM)

August 28, 2018

While the Cosmological Principle stipulate that ***the Universe is statistically homogeneous and isotropic***, we can easily imagine that statistics play a central role in cosmology. Indeed, the standard model assumes that all the perturbations in which galaxies and all the large scale structures were produced during inflation process. Since that, through the statistical properties (position, density, ...) of the matter we can evaluate the evolution of these primordial perturbations during the Universe history. For this reason, we decide to propose a lecture on cosmology mainly focused on the understanding of statistical observations and propositions. We will give 2 weeks of introductions on observational cosmology in order to be sure that all the basics are known and take advantage of this review to highlight the different statistical tools we will study in their context.

1 Program

1.1 Cosmology review lecture (8 h)

The idea of this first block is to review the observational cosmology basics. We recommend the following books for the students if they want to study more about cosmology: **???**

- 2h over the thermal history of the Universe
- 2h over the perturbations history
- 2h over Cosmic Microwave background perturbations (Temperature and Polarization)
- 2h over the large scale structures formation (BAO, clustering.....)

1.2 Boltzmann Solver and Basic tools (8h)

Before to start with the statistical tools, it is necessary to know how to code the basic observable for different cosmological models. First we need to implement the Hubble function and then use it on the 3 distances codes. We will play with various cosmological models (including curvature) in order to have a good intuition on the different evolutions.

Ina second time, we will start to use a Boltzmann solver code (CAMB or Class) which allows to generate most of the observable considering linear evolution of the perturbations from the inflation era up to the structure formation.

- 2h over personal Hubble function and distances codes.
- 2h lecture about Boltzmann codes and domain of validity.
- 4h over using CAMB (or Class) codes inside a python environment and generate the same functions and then generate the CMB statistical observables and the linear power spectrum of galaxies.

1.3 N-point statistics 8h

As we will see, most of the primordial statistical information is embedded inside the 2-point statistics (2-point correlation function and the Power Spectrum). However, the observations over galaxies have to take into account the non linear evolution of the perturbations implying that a lot of information move to higher order correlation functions. Moreover, the perturbations are dominated by the dark matter when we observe baryonic matter. So we also have to take into account the different evolution between these two quantities through the estimation of the bias.

- Probabilistic definition of the N-point correlation functions and the estimators. Introduction to the bias and variance of the estimator. (**??**).
- Code a brut force 2-point correlation function estimator and apply it on simulated data following different kind of underlying statistics.
- Generate random data following a Uniform/Poisson distribution.
- Apply (or code if enough time) the power spectrum on the same data. Introduction to the sample variance and shot noise.
- Code and play with a 3-point correlation function estimator, and its Fourier transform the bispectrum, on non gaussian data.
- Understand the link between 3-point statistics on CMB and the primordial non-gaussianities. Understand the link between the 3-point statistics on the galaxy distribution and the formation process of these objects and why it can be a good estimator for modified gravity probe.

1.4 Gaussian random field : Initial conditions and evolution (12h)

As we will see during the previous section, most of the information is still contained in the 2-point statistics justifying that the 2-point correlation function and its Fourier transform, the power spectrum, of the matter is of the first importance. A way to well understand these tools is to generate random data following the 2-point statistics information for various one. Moreover, the initial conditions for cosmological simulations need to use 3D Gaussian random field (and some distortions on top) before to run them. It is a crucial point because the cosmological principle stipulate that our Universe is one specific realization among an infinity of independent realization following the same underlying statistics. The only way we have to see if some observables are reproducible following a given statistics or not. A well known problem is the "cold spot" in the CMB temperature map.

- Generate 1D gaussian random trials for various power spectra. Study the variability on the results between different realization of a same random process. Then study the variability between different realizations following different power spectra.
- Generate 2D gaussian random trials. Do the same analysis than for the 1D case.
- Generate 2D gaussian trials projected on a sphere using spherical harmonics decomposition. That is a very powerful tool for the study of the CMB and galaxies distributions.

We will also introduce the way to evolve the density perturbations in different context:

- Low density contrast $\Rightarrow$ Linear theory
- Density contrast $\sim 1 \Rightarrow$ 2nd order perturbation theory (2LPT in our case) or simulations
- Density contrast $>> 1 \Rightarrow$ Simulations

1.5 Galaxy clusters: Mass Function and HOD (4h)

The most massive virialized structures in the Universe are the galaxy clusters. The mass probability distribution (average number of cluster expected by volume unit in a given range of mass and redshift) at different time of these objects is very sensitive on the presence of dark energy [**?**]. We can derive this mass function from the 2-point statistics of CMB assuming a spherical collapse model : the Press and Schechter formalism [**?**]. Using the power spectrum information at the recombination time, we can estimate the average number of over-densities in term of size and contrast density which are suppose to collapse at a given time. Knowing these two quantities, we can easily derive the mass of the formed clusters. However, we need to assume the linear theory of perturbation evolution to do this which lead to some dicsrepencies between the simulations and the Press and Schechter predictions. These discrepencies are corrected using semi-analitical models based on the Press and Schechter formalism modified adding some parameters fitted on high resolution simulations [**?**].

One time we obtain the halo mass function, we need to know the number of galaxies and their properties to rely the theory to the observation of the galaxies. The most important tool to do that is the Halo Occupation Distribution (HOD) [**?**]. The formalism allows to fit few parameters relied to a modified Press and Schechter function, the matter power spectrum and the number of galaxies we observed for a given survey. It means that we can model a survey through the type of galaxies, their luminosity/mass and the optical properties of the survey. The other important application for HOD is to populate cold dark matter simulations with galaxies. Indeed, most of the cosmological simulations are pure N-body cold dark matter particles. However, we do not observe the cold dark matter and we need to populate the simulations considering different existing/future galaxy surveys.

- Introduction to the spherical collapse model [?] and the Press and Schechter formalism.
- Generate 1D gaussian random trials following CMB power spectrum and vizualised in this toy model where the cluster will collapse in the future.
- Generate 2D gaussian random realization and do the same exercise. The goal is to deduce the form of the Halo Mass Function.
- Using real initial conditions of a simulation, redo the same work and compare with the halos we found at several time step of the simulation using the non N-body evolution
- Introduction to the HOD models and use it to populate the simulation for a BOSS/eBOSS survey.

1.6 Fitting methods and inference of parameters and errors (12h)

One of the main goals of the studies is to infer the best parameters of a model given data. Of course, best values is a small part of the game. The most important is to know the probability distribution of the posterior for the parameters.

1.6.1 χ^2, $\Delta\chi^2$ and Fisher Matrix (8h)

We will first focus on the well known χ^2 distribution and demonstrate this law. We will use this distribution in order to determine the p-value, so the goodness of fit, of simple models on generated data. Then we will see the $\Delta\chi^2$ method which allows to infer the parameters and the errors under the consideration of gaussian errors. We will draw the confidence contours and see the difference between using the Hessian matrix or to apply directly the $\Delta\chi^2$ criteria.

1.6.2 Bayesian inference of parameters and Monte Carlo Markov Chain (4h)

The previous method is generally associated with the frequentist approach (excepted the Fisher Matrix) in the sense that we do not use a prior information in order to infer the result. Due to the small number of independent measurement we can do in comparison to the number of free parameters of the standard model of cosmology, we have to take into account most of the data together. The standard way to do it is to use the bayesian inference where the previous likelihoods we get from other experiment are moved into the prior.

Another difference between χ^2 method and bayesian approach is the possibility to consider errors over the data points and the parameter estimation distinct to gaussian distribution.

However, this degree of freedom prevents to use a simple well define criteria to estimate the confidence contour. In order to evaluate them it is necessary to explore the parameter space in following a acceptation/rejection method for each step and finally consider the confidence contours as the limits of embedding a given probability presence.

- Introduction to Bayes theorem and use it on simple example
- Equivalence between χ^2 and Likelihood function in case of gaussian error.
- Introduction to Monte Carlo Markov Chain
- From the likelihood estimation to the posterior estimation through the prior in the step generator
- code a simple MCMC code and apply it on a simple example.
- use MontePython MCMC code and apply it on Supernovae data sample and estimate H_0 and Λ parameters.

Machine Learning (if time)

Very high level statistics as machine learning start to be used in cosmology, for the data analysis and compression as well as for the parameter inference [?; ?]. Indeed, Machine Learning is a large family of statistics which allows to use hidden variables in order to classify objects (can be time series, images....), infer parameters or take decisions. So, it also allows sometimes to do non-parametric analysis which can be powerful in some context but not every time. There are a huge number of different approaches taking into account:

- supervised, semi-supervised or unsupervised training sample
- methods Support Vector Machine, ABC, tree, random forest, neural network......

and each one needs different knowledges. Unfortunately, Machine Learning algorithms are often used as black-box which can lead to misleading conclusions. The main problem lies in the understanding of the noise and so to be able to evaluate the accuracy of the results. A Machine Learning can almost always give you a result but it does not mean that is a relevant one. A comparison could be to use the minimum χ^2 method without to compare the result with the theoretical distribution before to conclude about the best fit parameters.

So we propose to expose some methods of Machine Learning and to use them in context of simple examples.