A note on large scale structures

Pierros Ntelis,[a,1]

[a]Aix Marseille Univ, CNRS/IN2P3, Centre Physique de Particule a Marseille (CPPM), Marseille, France

E-mail: pntelis -at- cppm.in2p3.fr

DRAFT

Abstract. In this review, we present the basic concepts of physics on large scale structures. We start by giving a brief overview on the thermal history of our universe. We describe the theoretical framework, behind the magnificent scenery of observations, known as the standard ΛCDM model. The statistical observables currently used from surveys are discussed. We give an overview of the main limitation we have from observations. A summary of current and future large scale structure surveys is discussed. Finally, we give a brief overview on elements of statistics needed to study large scale structures physics.

DRAFT

Contents

DRAFT

1 A brief introduction

At first glance, we understand our universe as describe by Fig. 1. The universe starts from quantum fluctuations, for unknown reasons, as a hot dense plasma at about 13.8 Gyrs ago. The interaction between baryons, leptons, photons, and neutrini is violent. An inflationary growth of structure follows only after 10^{-32}sec after the "Big Bang". The photons are scatter from electrons and baryons and the universe is opaque. At about 380,000 yrs after the "Big Bang", the interactions of photons, baryons and leptons reduce rapidly as an epoch know as *Big Bang Nucleonsynthesis* (BBN). Therefore the photons are decoupled from the baryons and leptons and start to diffuse freely in space at an epoch know as *decoupling*. This radiation that is released we can observed it today and comes with the name *Cosmic Microwave Background* (CMB). At that time baryons and leptons combined with each other to form the first neutral hydrogen atoms at an epoch know as *recombination*, $z \sim 1090$. Due to the high transparency of structures, since baryons and electrons do not interact with the photons, photons were diffused rapidly. Therefore there is an epoch of our universe where matter does not interact with the light so we cannot observe it. Since we have no clue about that epoch we named it, *Dark Ages*.

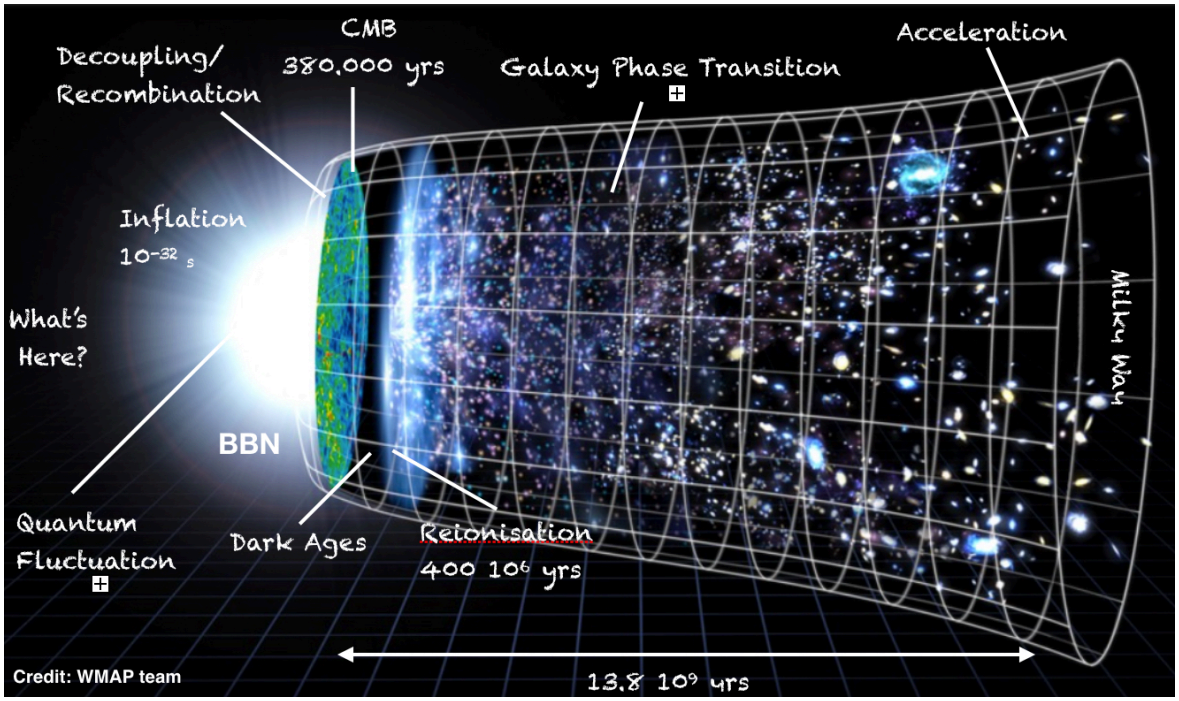

Figure 1. The current understanding of our universe. The universe starts as a hot densed fluid that expands through space for about 13.8 billion years. [See text for details]

About 400 million yrs after the "Big Bang" another phase change occurs, with the name *reionisation* epoch. Once objects started to condense in the early universe that were energetic enough to reionize neutral hydrogen. As these objects formed and radiated energy, the universe reverted from being neutral, to once again being an ionized plasma. This occurred between 150 million and one billion years after the Big Bang ($6 < z < 20$). At that time, however, matter had been diffused by the expansion of the universe, and the scattering interactions of photons and electrons were much less frequent than before the lepton-baryon

recombination. Thus, a universe full of low density ionized hydrogen will remain transparent, as is the case today.

At later times, the universe pass to the next phase transition, namely *structure formation*. At that time the first stars are forming. Those stars are primordial so they are made only by elements with atomic number as high as that of Iron, $^{56}_{26}$Fe. Due to the gravitational force of those massive objects and the expansion of the universe, these objects collide with each other. These collision produce high energetic phenomena, that allow the acceleration of particles. High speed moving particles produce collisions that produce the highest atomic elements such as that of Uranium, producing heavier stars, clusters of stars and in extend galaxies. Those collisions are happening to large gravitational potentials of another unknown matter of our universe which comes with the name *Dark Matter*. This exotic matter is responsible for the constant rotational curves of stars within a galaxy.

At the time of recombination and decoupling the luminous matter (baryons and leptons) and the dark matter interact with each other through gravity. At that time gravity is a attractive force. The space time is expanding. In order to compensate the expansion and the attractive force the hot dense fluid (baryons leptons and photons) is in an oscillatory phase. When the photons are released from matter and travel freely through space the baryonic fluid (baryons and leptons) is freezing in space. This effect is known as *Baryon Acoustic Oscillations* (BAO). The frozen modes are expanding with the comoving space through eternity. Finally the universe pass to each last phase which is the accelerated expansion, $z \sim 0.5$, due to a new aspect of gravity which acts as a repulsive force, currently with the name *Dark Energy*.

This magnificent scenery of phenomena are observed in an expanding background which was first observed by Hubble[1], in 1929, who studied the recessional velocities, v_{rec} of far galaxies and compare it to their distances from the earth, d, establishing their proportionality with a constant that we named after him, H_0, namely *Hubble constant*. This law is given by:

$$v_{rec} = H_0 d \ . \tag{1.1}$$

DRAFT

This law is still accurate today, in the approximation regime that was observed, and the Hubble constant has its best value, $H_0 = 67.27 \pm 0.66 \text{kms}^{-1}/\text{Mpc}$ as measured by the Planck Satellite[2] in 2015.

This is a summary of the standard model of modern cosmology which comes with the name ΛCDM. The "Λ" correspond to Dark Energy uknown component of our universe responsible for the accelarating nature of our universe, while the "CDM" stands for Cold Dark Matter, which is the most accepted feature of current dark matter models. This is only a brief introduction, therefore many simplifications were done for this discussion. The interested reader is directed to Dodelson [3] for a more detailed discussion.

2 An non exhaustive historical summary

The main theoretical framework of ΛCDM model was devellopped by an non excaustive list of fathers of modern cosmology, Einstein, Friedmann, Lemaître, Alpher, Gamow, Hubble, Zwicky, ... Einstein developed the Theory of General Relativity in 1915[4] which explains how gravity is only a manifestation of the effect of how the matter locally affects the curvature of space time around it. This Theory is formulated in the well know Einstein field equation

$$G_{\mu\nu}(x) = \kappa T_{\mu\nu}(x) \ , \tag{2.1}$$

where $G_{\mu\nu}$ is the Einstein tensor that describes the local curvature around a specific matter density which has a simplistic energy tensor, $T_{\mu\nu}$ and κ is a topological constant. In 1922, Friedmann[3] have applied the above formulation to a model for the evolution of the universe. He solved the above equations for a homogeneous and isotropic universe. At about the same time independently, in 1931, Lemaître have applyied the above equations for solutions of a homogeneous and isotropic universe that starts from a small hot dense region known as the "Big Bang Theory". Notice that this term is inaccurate since the equations break down on $r = 0$, an issue know as *singularity*. However this description agrees with the data that we have so far, which are not corresponding to $r = 0$ and therefore this model is still well accepted from the largest amount of the physics community, nowadays, 2018. This gave rise to the smooth cosmology that we are going to briefly describe in section 3.1. However, the standard model described by a "Big Bang" origin is still under investigation against the several attempts to explain some unexplained phenomena. Alternatives to this model are among the following: Late time isotropic universe, Bouncing universe, Multiverse, and so on.

Fast forward in time, 1948, Alpher a student of Gamow with his supervisor have predicted that the universe initially could emit a radiation at the Microwave regime known as the Cosmic Microwave Background (CMB) radiation. Therefore that year they published the so called α, β, γ-paper[5] introducing to the author list the friend of Gamow Bethe[1]. In 1965, Penzias and Wilson have observed by accident this radiation[6] without knowning the prediction of the Alpher and Gamow. Fast forward in time in 2006, Smooth and Mather get a Nobel Prize for the discovery of the Black Body nature of CMB by leading the team on observing with unprecedentedly accuracy the CMB temperature, $T_{CMB} = 2.728 \pm 0.004$ K as measured by the FIRAS instrument of COBE satellite[7]. This measurement confirmed the state of the art ΛCDM model which predicts the Baryon Acoustic Oscillation phenomenon behind the structure of the CMB radiation. Afterwards, WMAP confirmed this oscillatory feature by studying the temperature fluctuations of the CMB in 2010 [8]. Now the state of the art of this measurement is acquired by Planck satellite[2] measuring the CMB fluctuations unprecedentedly accuracy, $\Delta T/T \sim 10^{-5}$ at redshift, $z \sim 1090$ and inferring precision measurement on the parameters of ΛCDM model among alternatives.

In 1965, Gunn and Peterson observed the *Gunn-Peterson trough*, a feature of the spectra of quasars due to the presence of neutral hydrogen in the Intergalactic Medium (IGM) [9]. This marked the epoch of *reionisation* and it is a hot research area these days due to the lack of observations at $6 < z < 20$.

At the large scale structure observations there were several developments as well at that time. In 1929, Hubble[1] turned his telescope on the sky and observed the expansion of the universe confirming the theoretical framework build upon the previous authors. Hubble determined a linear relation between the recessional velocities (v_{rec}) and the radial distances of galaxies):

$$v_{rec} := cz = H_0 d \tag{2.2}$$

where z is the redshift, c is the speed of light and H_0 is the hubble expansion rate. The determination of the velocities was obtained by studying the spectrum of the galaxies . In order to obtained an estimate for their distance, he used Cepheids Variables. Cepheids are stellar objects within galaxies, that undergo pulsations in very regular periods on the order of days or months. This regular pulsations allows to determine the radial distance of the

[1] Bethe has nothing to do with this research but he was selected on the author's list by the authors only for the sake of the naming of the paper!

DRAFT

cepheids from the earth and in extend their host galaxies. The know diagram that Hubble built was named after him and it is shown in Fig. 2.

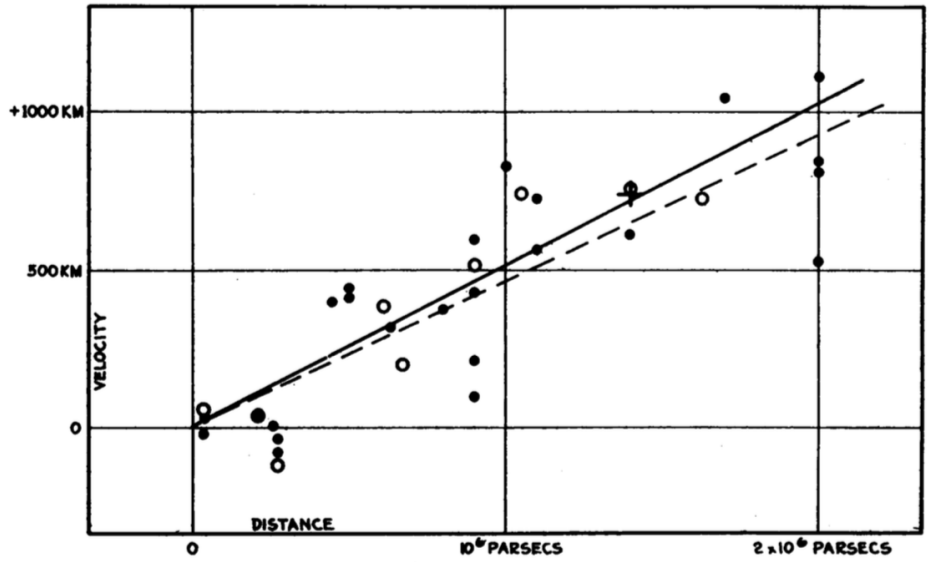

Figure 2. *Hubble diagram* obtained and the striking linear relation between the distances and the velocities as measured from Cepheids Variables by Hubble reveilling the expansion of our universe[1].

The *redshift*, z, is the shift of the emiting spectra (towards the red color) as observed by an observer on earth in respect of the rest frame emission of spectra as emitted by the galaxy itself. In other words, assuming that a galaxy emit radiation in a specific pattern (the spectrum which is the flux against different wavelengths), if this galaxy moves in respect to an observer away from him, it is nutural that the observed spectra will by shifted in respect of the emitted spectra due to dopler effect. This redshift is formulated as following:

$$1 + z = \frac{\lambda_{obs}(t_{obs})}{\lambda_{em}(t_e)} \ , \tag{2.3}$$

where $\lambda_{em}(t_e)$ is the wavelength of the emitted spectra at the rest frame of the galaxy at time t_e, while $\lambda_{obs}(t_o)$ is the wavelength of the observed spectra as observed by an observer at time t_o. We explain this phenomenon further in section 7.3. For more details the reader is redirected to [10].

In 1937, Zwicky [11] have studied the virial theorem,

$$(\text{Kinetic energy}) = -\frac{1}{2}(\text{Potential energy}) \ , \tag{2.4}$$

on galaxies in the local group and he realised that a missing compenents exists in order to much the observational data with the theoretical model of gravity, (Newtonian Gravity back then). The missing component was named Dark Matter since it was not luminous but was interacting with the luminous matter through gravity. This missing matter could compensate the missing mass that was necessary to fit the constant rotational curves at large radii away from the centers of the galaxies. A more robust confirmation of this phenomenon was aqcuired later on by Ruby[12] in 1970, who studied the rotation of the andromeda nebula from a spectroscopic survey.

DRAFT

Phenomenon	redshift, z	time, t	temperatue, T [eV]
Inflation	?	10^{-32}sec	?
Photon Decoupling	1090	380 k yr	$0.23 - 0.28$
Reionization	$11 - 30$	$100 - 400$ M yr	$2.6 - 7.0$
Accelerated Expansion	0.4	9 G yr	0.33 m
Present	0	13.8 G yr	0.24 m

Table 1. Key phenomena of large scale structures.

Fast forward in time, in 1998, three teams independently [13–15] have studied the luminosity distance relation of type Ia Supernovae and discovered the acceleration of the universe at late times $z \sim 0.4$, confirming once again the state of the art ΛCDM model. At 2005 indepedently two teams have observed through spectrscopic surveys the imprint of the Baryon Acoustic Oscillations at the distribution of galaxies another prediction of ΛCDM model. The one was the SDSS collaboration[16] and the other one was the 2dFGRS collaboration[17]. This discovery cannot be explain yet from alternatives of ΛCDM such as MOdified Newtonian Dynamics (MOND) theories[18]. For a more detailed discussion on alternative to ΛCDM theories the reader is redirected to Hamilton [19].

In the standard model of cosmology, redshift, time and temperature are related quantities. We summarise the discussed key phenomena of large scale structure in those parameters in table 1. You can see that some boxes are still missing particular in the inflation paradigm which is still an active area of reasearch of CMB observations but this falls of the discussion of this manuscript. However there is still a vast amount of improvement on the understanding of our universe on large scales as well in order to complete this picture. The BAO is the main observable in the late universe on large scale structures and we hope to give an overview on the rest of this document and we hope to reveal to the reader the exciting discoveries that await the large scale structures observations.

DRAFT

3 Theoretical Framework

The basic theoretical framework is build upon general theory of relativity and quantum field theory. Semianalytical models of the latter framework, alongside with observational evidence led the physics community to build the ΛCDM model, as we explained before. Now we will briefly describe the "smooth" part of this model and the "perturbed". To describe this framework, we will use information from [3, 10, 20]. For a more complete description please read the aforementioned references.

3.1 Smooth Cosmology

Friedmann[2] has shown in the 20's that one can use the cosmological principle to build such a coordinate system in order to solve Einstein Field Equations for a dynamical model that describe an expanding, homogeneous and isotropic universe, or *smooth universe*. There he showed that for a statistically homogeneous and isotropic universe, an observer has a coordi-

[2]Lemaître, Robertson and Walker, independently of Friedmann, have developed the same model during same epoch, 1920-1930.

nate system that follows the FLRW metric[3]:

$$ds^2 = -c^2dt^2 + a^2(t)\left[\frac{1}{1-kr}dr^2 + r^2(d\theta^2 + sin^2\theta d\phi^2)\right] \tag{3.1}$$

where $d\Omega = d\theta^2 + sin^2\theta d\phi^2$ reflects the isotropy condition and $\gamma_{ij} = \frac{1}{1-kr^2}dr^2 + r^2 d\Omega$. If $k = 0$, space is flat and infinite (critical). If $0 < k < 1$ space is spherical and finite (closed), while $-1 < k < 0$ correspond to a hyperbolic and infinite space (open).

The FLRW metric is used as an input on the left hand side of equation Eq. 2.1 for the computation of the scale factor as a function of the geometrical properties of the universe. Thus one can easily show that the Einstein Tensor, for an FLRW metric (Eq. 3.1), reduces to a tensor with the following non-zero components:

$$G_{00} = 3\left(\left[\frac{\dot{a}(t)}{a(t)}\right]^2 + \frac{kc^2}{a^2(t)}\right), \qquad G_{ij} = -\gamma_{ij}\left[k + 2a(t)\ddot{a}(t) + \dot{a}^2(t)\right] \tag{3.2}$$

where dot " $\dot{}$ " represent the derivative in respect of time t.

The right hand side of equation Eq. 2.1 describes the energy content of the universe as a perfect fluid in thermodynamic equilibrium, thus the energy-stress tensor takes the simplified form:

$$T_{\mu\nu} = \left[\rho(x) + \frac{P(x)}{c^2}\right]u_\mu u_\nu + P(x)g_{\mu\nu} \tag{3.3}$$

where $\rho(x)$ is the energy density, $P(x)$ is the pressure and u_μ is the 4-velocity. The cosmological principle implies that $u^\mu = (1, 0, 0, 0)$, meaning that the fluid is locally at rest with respect to the chosen frame. Furthermore, the cosmological principle restricts the energy density and pressure to be constant over space but allows a possible time dependence. These considerations model the stress energy tensor, with only non-zero components, as follows:

$$T_{00} = \rho(t), \qquad T_{ij} = \frac{P(t)}{c^2}\, a^2(t)\gamma_{ij} \tag{3.4}$$

DRAFT

The gauge invariance allows us to add a constant on the Eq. 2.1, *cosmological constant*, Λ. By taking all the above considerations into account, the 00-component and the trace of Eq. 2.1 are written as:

$$H^2(t) \equiv \left(\frac{\dot{a}}{a}\right) = \frac{8\pi G}{3}\rho(t) - \frac{kc^2}{a^2(t)} + \frac{\Lambda c^2}{3} \tag{3.5}$$

$$-\left(\frac{\ddot{a}}{a}\right) = \frac{8\pi G}{2}\left[\rho(t) + \frac{3P(t)}{c^2}\right] - \frac{\Lambda c^2}{3} \tag{3.6}$$

where the $H(t) = \dot{a}/a$ is the Hubble expansion rate. The above differential equations are not enough to completely specify the system, i.e. $a(t), \rho(t)$ and $P(t)$. Thus, either by combining the above equations or by using the local conservation of the stress-energy tensor ($T^{\mu\nu}{}_{;\mu} = 0$), we have that:

$$\dot{\rho}(t) = -3H(t)\left[\rho(t) + \frac{P(t)}{c^2}\right] \tag{3.7}$$

The set of the 3 latter equations (Eq. 3.5, Eq. 3.6 and Eq. 3.7) are used to describe the evolution $a(t)$ of the cosmic fluid with properties $\rho(t)$ and $p(t)$. This set of equations are

[3]For an approach that has more general considerations of topology on constructing this metric, see Appendix A2: Topological Restrictions from Ntelis [10]

called *Friedmann* equations. However, in the ΛCDM-modelling there are several species of the total cosmic fluid such as $X = \{\gamma, \nu, b, cdm, \Lambda\}$ which corresponds to photons, baryons, neutrinos, cold dark matter and dark energy, respectively.

The cosmic species are divided into two general categories, i.e. the *relativistic* and *non relativistic* species, according to the level of their rest mass energy mc^2. The former have a rest mass energy which is insignificant against their average kinetic energy $mc^2 << k_B T$. This leads to $P_{rel} = \rho_{rel}/3$. The latter are those whose momentum is negligible to their rest energy ($mc^2 >> k_B T$), and therefore $P_{n.rel} \simeq 0$. However, one may generalise those two approximated relations for the two categories of species with a parameter

$$w = \frac{P}{\rho c^2} \tag{3.8}$$

namely *equation of state* parameter. This allow for a class of solutions of Eq. 3.7, i.e.

$$\rho_X(t) \propto [a(t)]^{-3(w_X+1)} \tag{3.9}$$

for each species X.

It is convenient, now, to define the critical energy density as the energy density for a universe of zero curvature (k=0) and no cosmological constant ($\Lambda = 0$):

$$\rho_c(t) = \frac{3H^2(t)}{8\pi G} \tag{3.10}$$

Then by dividing Eq. 3.5 with $H^2(t)$, we have:

$$1 = \frac{8\pi G}{3H^2(t)}\rho(t) - \frac{kc^2 8\pi G}{8\pi G a^2(t) H^2(t)} + \frac{\Lambda c^2 8\pi G}{3 \times 8\pi G H^2(t)} \tag{3.11}$$

Now substituting Eq. 3.10 to Eq. 3.11 we have:

$$1 = \frac{\rho(t)}{\rho_c(t)} - \frac{1}{\rho_c(t)}\frac{kc^2}{8\pi G a^2(t)} + \frac{1}{\rho_c(t)}\frac{\Lambda c^2}{3 \times 8\pi G} \tag{3.12}$$

Last but not least, we introduce the ratio of energy densities of the possible species (X) of our universe against the critical density as:

$$\Omega_X(t) = \frac{\rho_X(t)}{\rho_c(t)} \tag{3.13}$$

where $X = \{\gamma, \nu, b, cdm, \Lambda\}$ correspond to photons, baryons, neutrinos, cold dark matter and dark energy, respectively. One may define as well the energy density ratio of curvature as:

$$\Omega_k(t) = -\frac{kc^2}{8\pi G a^2(t)} \ . \tag{3.14}$$

Therefore, all those species must satisfy the local energy conservation equation at all times:

$$\Omega_k(t) + \sum_X \Omega_X(t) = 1 \ . \tag{3.15}$$

Thus in the field of concordance cosmology, we use the above simple parametrization (Eq. 3.13) to measure the ratio of energy densities of the different species in our universe. The convention we adopted is that when we drop the time dependence, we talk about the energy density ratio today $\Omega_X = \Omega_X(t = 0)$.

DRAFT

3.2 Perturbed Cosmology

The latest observations, such as the CMB measurements[21], have shown that the universe is full of small inhomogeneities. In other worlds the universe behaves not homogeneously at smaller scales. Two ingredients are necessary to describe the unforementions inhomogeneities. The first ingredient is the perturbation of the smooth metric, i.e. FLRW metric (Eq. 3.1). The second one is the Boltzmann equation that describe the nature of the interactions and the evolution between the different species of the cosmic fluid, beyond the equilibrium.

In order to account those inhomogeneity in the previous described framework (section 3.1), we perform a perturbations in the metric. The mechanism of perturbation give rise to a gauge-invariance. In our case, we are going to describe the simplest one, the Newtonian synchronous gauge. The FLRW metric in the perturbation theory is going to be written as:

$$ds^2 = -\left[1 + 2\Psi(\vec{x})\right] dt^2 + a^2(t)\left[1 + 2\Phi(t)\right] d\vec{x}^2 \tag{3.16}$$

where we have considered scalar perturbations defined via the $\Phi(t)$ spatial curvature field and the $\Psi(\vec{x})$ Newtonian potential field. By neglecting Ψ and Φ scalar perturbations, we retrieve the homogeneous and isotropic, FLRW metric.

The Perturbed Boltzmann Einstein Equations can be summarised by

$$\mathcal{D}_t f_X(\vec{x}, \vec{p}, t) = \mathcal{C}[f_X(\vec{x}, \vec{p}, t)] \tag{3.17}$$

where the left hand side describes the time evolution of the distribution $f_X(\vec{x}, \vec{p}, t)$ of the primordial fluctuations of each species, which we have developed in first order approximation. Note that in first order approximation we have that:

$$\mathcal{D}_t = \partial_t + a^{-1}(t)\hat{p}^i + \partial_t\Phi(t) + a^{-1}(t)\hat{p}^i\partial_i\Psi(\vec{x}) \tag{3.18}$$

DRAFT

which is the well defined derivative of the perturbed metric. The right hand side of Eq. 3.17 describes the collision treatment between the different species X, $C[f_X]$. For the interaction between photons and leptons, we consider the classical *Thomson scattering* non-relativistic approach, $l^{\mp} + \gamma \leftrightarrow l^{\mp} + \gamma$ with an interaction rate $\Gamma \simeq n_l\sigma_T$, where $\sigma_T \simeq 2 \times 10^{-3} MeV^{-2}$ is the Thomson cross section.

For cold dark matter, we consider a collisionless non-relativistic approach, as done in various famous structure formation history models. This are the simplest models that agree with observational large scale structure data. For baryons and leptons interactions, we assume a *Coulomb Scattering*, $b^{\pm} + l^{\mp} \leftrightarrow b^{\pm} + l^{\mp}$ in the Quantum ElectroDynamic (QED) approach. While for neutrini, we only consider them as a massless relativistic particle fluctuation overdensity and therefore we assume that they do not interact with matter. This is true only in the linear regime at large scales. Adopting a Fourier transform framework to simplify the equations in question, we end up with a set of 6 linear differential equations describing the non linear evolution of the 3 different species of density fluctuations (baryons, photons and neutrinos and Dark Matter) and their corresponding velocities at large scale as a function of conformal time[4], η, and wavenumber, $\vec{k}$. However, this system is coupled to the 2 degrees, $\Phi(\eta)$ & $\Psi(\vec{k})$, of freedom defined by the perturbations of the curved metric. Thus, in order to completely specify the system one may solve the time-time component and the spatial trace

[4] The conformal time $\eta = \int_0^t dt'/a(t')$ defines the time needed for particles that travel in the speed c to reach an observer from the maximum distance existing in the universe (*observable universe*) which we call *particle horizon*.

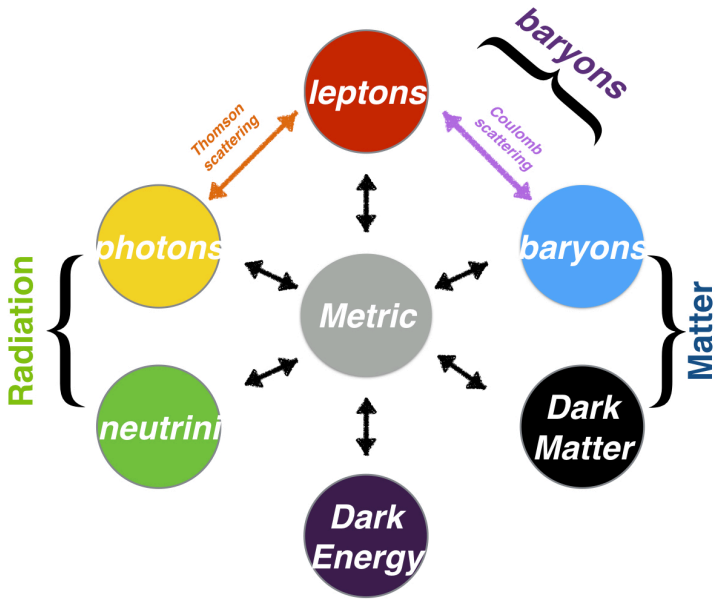

Figure 3. Schematic representation of the linear, coupled Boltzmann-Einstein field equation describing the interplay of physics at large scales, inspired by Fig 4.1 of Dodelson [3].

DRAFT

of the Einstein equations using the perturbed metric defined via Eq. 3.16. Thus we end up with the 8 coupled differential equations, namely *Perturbed Boltzmann-Einstein Equations* that completely specify the system on large scale structures, i.e. the evolution of the density and temperature fluctuations, $\delta_X(t,\vec{x})$ & $\frac{\delta T}{T}|_X(t,\vec{x})$, of the different species X. This interplay between the different species and the metric is represented schematically in Fig. 3.

$\vec{\Omega}$	Value (Planck 2015)	Physical Description
$\omega_{cdm} = \Omega_{cdm}h^2$	0.1198 ± 0.0015	physical cold dark matter density ratio
$\omega_b = \Omega_b h^2$	0.0225 ± 0.0002	physical baryon density ratio
$h = H_0/100$[Km/s/Mpc]	0.6727 ± 0.0066	dimensionless hubble expansion rate
n_s	0.9645 ± 0.0049	spectral index
$\ln[10^{10}A_s]$	3.094 ± 0.0034	Amplitude of the primordial fluctuations
Ω_m	0.316 ± 0.013	Total matter density ratio
Ω_Λ	0.684 ± 0.013	Dark Energy density ratio
Ω_k	-0.004 ± 0.0015	curvature density ratio
w_0	-1.006 ± 0.0045	equation of state parameter today
w_a	-0.0001 ± 0.0005	equation of state parameter redshifted

Table 2. Top: Standard parametrization of the ΛCDM model parameters as they could observed now, $z = 0$. Middle: Derived parameters Bottom: Extensions.

The full set of those differential equations are given in Chapter 4 equations 4.100-4.107 of Dodelson [3].

3.3 Theoretical prediction

All the above description encodes the total theoretical framework that we confront in observations. The aforementioned Perturbed Boltzmann Einstein equations are solved semi analytically to give a prediction of our universe. This predictions is encoded mostly to the Power Spectrum of the total matter of the universe P_{matter} as a function of the wave-vector k, time, defined by z, and the cosmology dependence, $\vec{\Omega}$, as:

$$P_{matter}(k, z; \vec{\Omega}) = A_s k^{n_s - 1} T^2(k; \vec{\Omega}) D^2(z) \tag{3.19}$$

These equations describe the intrinsic power shape of the matter density field coming from the solutions of the Boltzmann-Einstein equations. $T(k; \vec{\Omega})$ is the transfer function[3, 22] which is basically how the modes of the primordial fluctuations of the total matter field in our universe before the last scattering surface are converted to the primordial fluctuations of the total matter field of our universe after the last scattering surface, which is denoted as the so called *drag epoch*. The $A_s k^{n_s - 1}$ factor describe the primordial shape of the power spectrum, where A_s the amplitude and n_s the spectral index of the primordial scalar power spectrum. Finally, $D(z) = G(z)/G(z = 0)$ is the *normalized scale independent linear growth-factor* which describes the evolution of the shape of the matter density field at different times according to redshift, z.

The prediction is always under development and there are several working package that allow one to obtain the numerical solution of the matter field. Currently the most popular ones are the CLASS[23] and CAMB[24]. These functionals are usually parametrized in the standard ΛCDM model by the parameters $\vec{\Omega}$ in the standard framework, or extensions of it, as shown table 2.

The standard parametrization includes; $\omega_{cdm} = \Omega_{cdm} h^2$ and $\omega_b = \Omega_b h^2$ the physical density of cold dark matter and baryonic matter, $h = H_0/100$[Km/s/Mpc] the dimensionless hubble expansion rate, n_s and $\ln[10^{10} A_s]$ the spectral index and amplitude of the primordial fluctuations. Derived parameters are usually the total matter density ratio, Ω_m, and the total Dark Energy density ratio, Ω_Λ. However there is the possibility for extensions such as for a variation of the curvature density of the universe, Ω_k. Other extensions include varying equation of state, w. A popular redshift dependence parametrizsation of the equation of state is given by

$$w(z) = w_0 + \frac{z}{1 + z} w_a \tag{3.20}$$

where w_0 is the equation of state parameter today and w_a is the difference between the equation of state parameter at high redshifts, usually $z >> 1$. This is the target of several experiments, that try to investigate the nature of Dark Energy field.

4 What is a structure?

Since we have describe the underlying physics of large scale structure, in this section we are going to answer to two relevant questions: what is a structure and what are the observables that we use on large scale structures. To answer to the first question let's review what are the structures observed so far. From small scales to large scales we can have different kind of structures and therefore different observables. We depict some of the main structures in Fig. 4. Let's start at the smallest possible scales, $l_{pl} \simeq 10^{-35}$ m, where the matter hypothetized to be structured as *strings* and/or as the *quantum foam*[10]. At scales $\delta l \simeq (10^{-20}, 10^{-15})$ m,

DRAFT

the matter is structured as *field particles* such as quarks, leptons, photons, W and Z bosons, Higgs and many more. At scales up to 10^{-10} m, the matter is structured in the form of *atoms* such as the form Hydrogen, and *molecules* such the ones of the water we drink, H_2O. These scales are usually described by the Standard Model build upon, Quantum Field Theory and/or Quantum Gravity models.

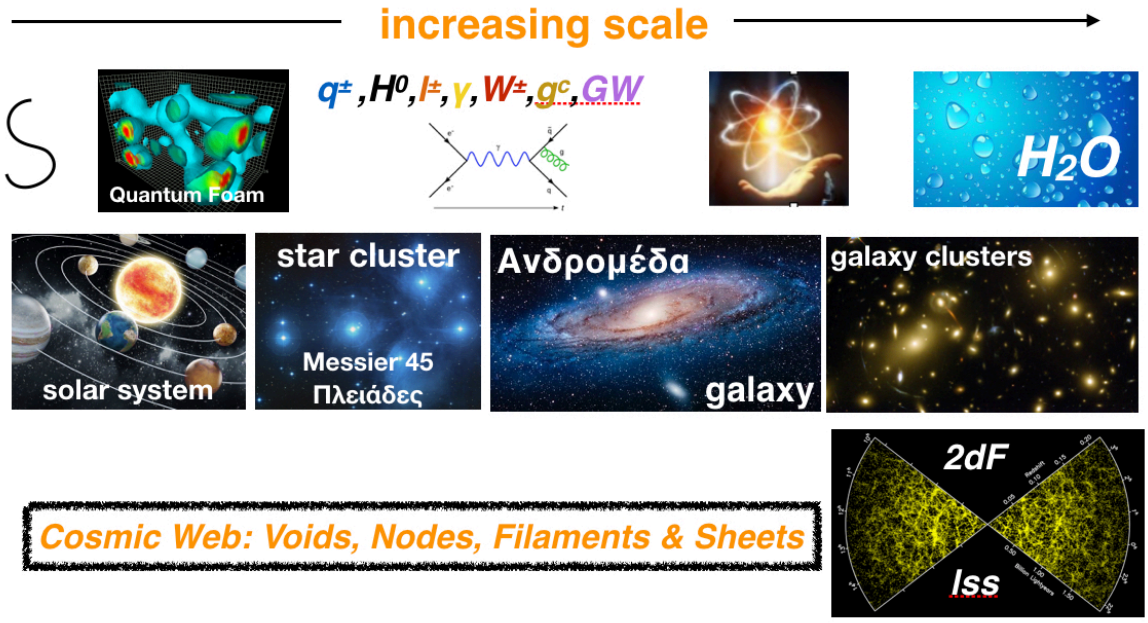

Figure 4. Illustration of the different matter structures in our universe in terms of scales, from the smallest hypothetized scales, $l_{pl} \simeq 10^{-35}$m, to the largest observable scales, of the order of Gpc$\simeq 10^{26}m$. [See text for details.]

DRAFT

At scales between $10^{-4}-10^{8}$m are the trivial scales that we make some easy observations, such as the ants walking in the surface of our the planet, we leave in!

At scales of the order of Astronomical Units (A.U.), which is the distance between the earth and our sun, 1 A.U. $= 10^9$ m, the matter is structure in the form of planets that travel around a massive object such as a *star*. This groups are called *solar systems*. Going further upwards on scales, the matter is structured in the form of *star clusters* such as the Messier 45 which in greek is called *Pleiades*, which in free translation means many.

On large scales of the order of kpc$(= 10^{19}$m) the matter is structured in the form of *galaxies*, *quasars*, *Super Novae* (SN). Galaxies are collection of stars which are concentrated in deep gravitational potentials. Quasars or Quasi Stellar Objects (QSO) are the most massive galaxies in the early universe. Explosions of stars or galaxies in the late time universe. Further classification includes galaxy clusters that are collection of galaxies in large gravitational potential at the order of few kpc. Do not forget that also there is the *dark matter* particles that usually are invisible to our instruments but they are being traced by the observing the dynamics of the luminous matter such as the galaxies. At the largest possibly observables scales, order of few Gpc$(\simeq 10^{26}$m), there are also some other classification of the large scale structure of the universe, in terms of *Nodes*, *Filaments*, *Sheets* and *Voids*, namely *Cosmic Web*. This form of structure can be also be consider as structure of foams. Therefore one can observe that at the very small scales, infinitesimally small and at the very large scales, infinitesimally

large , the universe is structure as a foam, or namely *large scale structure foam*! If one would like to describe marninally, the two structures towards the two infinities, he would give them a proper name, namely the *cosmic foam*. The *cosmic foam* is the form of structure at the infinitesimally small and the infinitesimally large scales of our universe. In other words, one would say that the universe seem blurry at the two infinities! How we can optimally combine these observations to distinguish between models is still an open question of current research!

The second question, what is our observable, is more complicated to answer. As you understood a structure can have different forms, and we can pick different ways to model it and study it according to our preference and convinience. On larger scale structures, we are interested on how matter is distributed. Therefore we are interested in the light coming from distant objects such as galaxies and the early universe (CMB radiation) which are able trace the total matter fluid of our universe. Therefore radiation in the form of photons is the main observable that allows us to observe the largest possible scales. Additionally, we are going to treat galaxies as point source of photon radiation to study their statistical properties.

These objects helps to constrain our theoretical models of cosmology such as the standard ΛCDM model or modifications of gravity and several submodels of this universal model such as galaxy formation models, Dark Energy models, Dark Matter models etc.

5 Baryon Acoustic Oscillation: Briefly

According to the short summary of our standard ΛCDM model, now we are able to give the insights of the famous phenomenon, Baryon Acoustic Oscillations.

About 300,000yr after Big Bang, the universe was in a hot and dense state expanding rapidly. Baryons , leptons and photons were interacting with each other due to high temperature. Note that the gravitational potential, which is created by the total matter, is attractive. While the kinetic energy coming from the high temperature and interaction of the particles was high. This creates an outward pressure on the structures. These counteracting forces of gravity and pressure created oscillations in the structure of total matter of the universe, analogous to sound waves created in the air by pressure differences.

DRAFT

At about $\sim 360,000$yr after the Big Bang (Recombination epoch), the universe cool down at a point were the baryons are combined with the leptons. These acoustic oscillations freeze.

Shortly after, about $\sim 380,000$yr after Big Bang (Decoupling epoch), the photons cannot interact anymore with the already produced atoms. Therefore the photons free stream in space-time continuum, in the form of CMB radiation. Note the informations of BAO oscillations is also encoded to the free streaming photons.

For about 13Gy the structures are evolving and therefore, the frozen Baryon Acoustic Oscillations evolve as well.

About $\sim$ 13Gyr after the Big Bang, we observe these frozen BAO in the distant universe! We observe them e;ither in the form of temperature fluctuations (CMB) or in the form of density fluctuations (galaxy distributions).

Note that this is only a short simplification of the phenomenon.

6 Observations: Clustering statistics

Observations are usually devided into two big categories; the primordial observations and the late time observations. In the first kind of observations we study the "initial" conditions in

the far past, $z \simeq 1100$ through local temperature fluctuations, $\delta_T(x)$ in the comoving position x. These are usually referred to as *CMB experiment*[2]. In the second kind of observations we study the "final" conditions of our universe in the near past, $z \simeq 1$, using the galaxy number density fluctuations, $\delta(x)$. These are usually referred to as *redshift surveys* or *galaxy surveys*. Both are used individually or in combination to constrain cosmological models. Let's focus on the redshift surveys.

6.1 The number overdensity field

Any redshift survey will observe a particular "window" of the universe, consisting of an angular mask of the area observed, and a radial distribution of galaxies. In order to correct for a spatially varying galaxy selection function, we translate the observed local galaxy number density, $n(t, x)$, as:

$$\delta(t, x) = \frac{n(t, x) - \bar{n}(t)}{\bar{n}(t)} \tag{6.1}$$

where $\bar{n}(t)$ is the expected mean density at a given time, t. At early times, or on large-scales, $\delta(t, x)$ has a distribution that is close to Gaussian one[2] and thus the statistical distribution is completely described by the two-point functions of this field.

6.2 The two point correlation function

The two point correlation function is the expected 2-point function of the aforementioned statistic:

$$\xi(t, x_1, x_2) \equiv \langle \delta(t, x_1)\delta(t, x_2) \rangle \ , \tag{6.2}$$

where the operator $\langle\rangle$ denotes average over space, $x = x_1 + x_2$. The *ergodic theorem* [5] that the space average is equivalent of the average over different realizations of the galaxy distribution.

From statistical homogeneity and isotropy, we have that:

$$\xi(t, x_1, x_2) = \xi(t, x_1 - x_2) = \xi(t, |x_1 - x_2|) \ . \tag{6.3}$$

An alternative definition of the correlation function, will help us understand better its statistical inference. Assume that we have two small regions in comoving space, δV_1 and δV_2, separated by a distance r. Then the expected number of galaxies, dP_{pair}, with one galaxy in δV_1 and the other in δV_2 is given by:

$$dP_{pair} = \bar{n}^2 \left[1 + \xi(r)\right] \delta V_1 \delta V_2 \ , \tag{6.4}$$

where $\bar{n}$ is the mean number of galaxies per unit volume. Therefore, $\xi(r)$ measures the excess probability of finding a galaxy at a finite volume separated by the referenced one by a given distance, r.

For $\xi(r) = 0$, the galaxies are unclustered (randomly distributed) on this scale - the number of pairs is just the expected number of galaxies in δV_1 times the expected number in δV_2. The values of $\xi(r) > 0$ correspond to strong clustering and $\xi(r) < 0$ to anti-clustering. The estimation of $\xi(r)$ from a sample of galaxies will be discussed in section 6.8.

[5] Note: The *ergodic theorem* suggest that an ensemble spatial average (average over many realisations) for a field, is equal to the spatial average over one realization, only if the field is random and large enough. In other words, all the information that is present in a complete distribution $p[\psi(\vec{x})]$ is encoded from a single sample $\psi(\vec{x})$ over all space. In other words, this means that spatial correlations decay sufficiently rapidly with separation such that many statistically independent volumes exist in one realization.

DRAFT

6.3 The power spectrum

Another convenient tool to measure the clustering in Fourier space is the power spectrum. The following Fourier transform is adopted:

$$\delta(k) = \int d^3r \delta(r) e^{ikr} \ , \tag{6.5}$$

$$\delta(r) = \int \frac{d^3r}{(2\pi)^3} \delta(k) e^{-ikr} \ . \tag{6.6}$$

The power spectrum is defined via:

$$P(k_1, k_2) = \frac{1}{(2\pi)^3} \langle \delta(k_1) \delta(k_2) \rangle \ , \tag{6.7}$$

with statistical homogeneity and isotropy giving:

$$P(k_1, k_2) = \delta_D(k_1 - k_2) P(k_1) \tag{6.8}$$

where δ_D is the Dirac delta function of a 3D field. The correlation function and the power spectrum for a Fourier pair as:

$$P(k) = \int \xi(r) e^{ikr} d^3r \ , \tag{6.9}$$

$$\xi(k) = \int P(k) e^{-ikr} \frac{d^3r}{(2\pi)^3} \ , \tag{6.10}$$

since they provide the same information. The choice of which to use is therefore somewhat arbitrary, see discussion by Percival [25].

DRAFT

6.4 Fractal Dimension

Nature all around us is full of fractal patterns, be it from amazing snowflakes to romanesco broccoli. Naturally due to the peculiar structure of galaxies around us, there were several claims that the universe behaving like a fractal one (Labini et al. [26], Coleman and Pietronero [27]). Therefore we can study the fractality of the galaxy distributions using the *fractal dimension*, defined as:

$$D_2(r) \equiv \frac{d \ln N(< r)}{d \ln r} \tag{6.11}$$

where $N(< r)$ is the *count-in-spheres* of galaxies. However the information obtained from this statistics it is similar to the one of $\xi(r)$ or $P(k)$ but it is an active field of research on how we can use it to optimize our observations! It has been shown that the universe at small scales behaves as a fractal, at scales less than $50h^{-1}$Mpc and at large scales it reaches a statistically homogeneous behaviour asymptotically[28–30]. For a homogeneous distribution the counts-in-spheres scale as the volume, $N(< r) \propto r^3$, while for a fractal distribution they scale as $N(< r) \propto r^{D_2}$ where D_2 quantifies the fractality of the distribution in study. For further discussion, see [10] and references therein.

6.5 Anisotropic statistics

Although the true universe is expected to be statistically homogeneous and isotropic, the observed one is not so due to a number of observational effects discussed in section 7. Statistically, these effect are symmetric around the line-of-sight. In the distant-observer limit these effects possess a reflectional symmetry along the line-of-sight looking outwards or inwards. Thus, to first order, the anisotropies in the over-density field can be written as a function of μ which is the cosine angle to the line-of-sight. Consequently, we often write the correlation function $\xi(r,\mu)$, the power spectrum $P(k,)$ or the Fractal Dimension, $D_2(r,\mu)$. It is common to expand these observables in Legendre polynomials $L_l(\mu)$ as:

$$\xi(r,\mu) = \sum_l \xi(r) L_l(\mu) \ . \tag{6.12}$$

Only the first thre even Legendre polynomials are important, as shown by Kaiser [31]; $L_0(\mu) = 1$, $L_1(\mu) = (3\mu^2 - 1)/2$ and $L_4(\mu) = (35\mu^4 - 30\mu^2 + 3)/8$. In the absence of redshift space distortions, discussed in section 7.3, only the "monopole" or the andgle-averaged correlation function survives:

$$\xi_0(r) = \frac{1}{2}\int_{-1}^{1} d\mu \ \xi(r,\mu) \ . \tag{6.13}$$

In the same way, one can expand the fractal dimension in terms of the legendre polynomials:

$$D_2(r,\mu) = \sum_l D_2(r) L_l(\mu) \ . \tag{6.14}$$

and study the anisotropic fractality of large scale structures. Another active area of current research[10]!

DRAFT

6.6 Higher order statistics

At early times and on large scales, we expect the over-density field to have Gaussian statistics. This follows from the central limit theorem, which implies that a density distribution is asymptotically Gaussian in the limit where the density results from the average of many independent processes. The over-density field has zero mean by definition so, in this regime, is completely characterised by either the correlation function or the power spectrum. Consequently, measuring either the correlation function or the power spectrum provides a statistically complete description of the field. To capture them on small scale structure, we usually resort to higher order statistics of the matter field. Higher order statistics tell us about the break-down of the linear regime, showing how the gravitational build-up of structures occurs and allowing tests of General Relativity.

The extension of the 2-pt statistics, the power spectrum and the correlation function, to higher orders is straightforward. From Eq. 6.4 we have:

$$dP_{tuple} = \bar{n}^n \left[1 + \xi^{(n)}\right] \delta V_1 \dots \delta V_n \tag{6.15}$$

The immediate application of n-point statistics is the *Bispectrum*, $B(k_1,k_2)$ defined as:

$$< \delta(k_1)\delta(k_2)\delta(k_3) >= (2\pi)^3 B(k_1,k_2)\delta_D(k_1 - k_2 - k_3) \tag{6.16}$$

studying the 3rd order statistics, and the *Trispectrum*, $T(k_1,k_2,k_3)$, defined as:

$$< \delta(k_1)\delta(k_2)\delta(k_3)\delta(k_4) >= (2\pi)^3 T(k_1,k_2,k_3)\delta_D(k_1 - k_2 - k_3 - k_4) \tag{6.17}$$

studying the 4th order statistics. As an example take a look at [32–35]. We study these higher order statistics to obtain information about the non-linear properties of the matter field which is caused by the small scale complex astrophysical processes, $r \sim 10\ h^{-1}$Mpc.

6.7 Non-Gaussianity

At very large scales the power spectrum is theoritized to deviate from the Gaussian case. It has been shown that primordial non-Gaussianities are generated in the conventional scenario of inflation[36]. However there are very weak. The primordial non-Gaussian properties[36] of the matter field are model in many different ways and can also be observed by different kind of observables. In our case we are going to focus to the power spectrum observable and the so called "local" non-Gaussianity. The "local" model is given by:

$$\Phi = \phi + f_{nl}(\phi^2 - \langle\phi^2\rangle)\ . \tag{6.18}$$

Here ϕ denotes the Gaussian random field while Φ denotes the Bardeen's gauge-invariant potential, which on sub-Hubble scales reduces to the usual Newtonian peculiar gravitational potential, up to a minus sign. On even larger scales this potential is related to the conserved variable ζ by

$$\zeta = \frac{5+3w}{3+3w}\Phi \tag{6.19}$$

where w is the equation of state of the dominant component in the universe. The amount of primordial non-Gaussianity is quantified by the non-linearity parameter $f_{\rm NL}$. Note that, since $\Phi \simeq \phi \simeq 10^{-5}$ then $f_{\rm NL} \sim 100$ which corresponds to relative non-Gaussian corrections of the order of 10^{-3}. While ζ is constant on large scales, Φ is not. Thus, there are usually two conventions for Eq. 6.18. The large scale structures (LSS) and the cosmic microwave background (CMB) one. In the LSS convention, Φ is linearly extrapolated now, $z = 0$. In the CMB convention Φ describes the primordial non Gaussian potential. Therefore, there is an approximate formula that relates the two observables:

$$f_{\rm NL}^{\rm LSS} = \frac{G(z=\infty)}{G(z=0)} f_{\rm NL}^{\rm CMB} \sim 1.3 f_{\rm NL}^{\rm CMB} \tag{6.20}$$

where $G(z)$ denotes the linear growth suppression factor relative to the Einstein-de Sitter universe. It is customary to report valus of $f_{\rm NL}^{\rm CMB}$ but there is also convient for the large scale structure community to use $f_{\rm NL}^{\rm LSS}$.

So far observations have shown that f_{NL} measurements are consistent with 0, and therefore no detection of deviations from the Gaussianity of the matter field were observed yet. However, this is another active field of research[37]!

6.8 Estimators for Correlation function and Fractal Dimension

Let's brefily discuss the estimators of the correlation function and the fractal dimension. For the correlation dimension what we can observe is the counts-in-cells. These are usually denoted as pairs of galaxies in different cells of r on a given galaxy data catalogue, $GG(r)$. Therefore we can formally define:

- $gg(r) = \dfrac{GG(r)}{n_g(n_g-1)/2}$, the normalized number of galaxy pairs separated by r,
- $rr(r) = \dfrac{RR(r)}{n_r(n_r-1)/2}$, the normalized number of random-point pairs separated by r,

DRAFT

- $gr(r) = \dfrac{GR(r)}{n_g n_r}$, the normalized number of galaxy random-point pairs separated by r,

where n_g and n_r are the total number of galaxies and random points, respectively. A usual estimator of the correlation function is the Peebles-Hauser estimator, $\widehat{\xi}(r) = dd(r)/rr(r) - 1$, for the two-point correlation function Peebles and Hauser [38]. This estimator is known to be less efficient than the more sophisticated Landy and Szalay [39] estimator.

$$\widehat{\xi}_{ls}(r) = \frac{gg(r) - 2gr(r) + rr(r)}{rr(r)} \,, \tag{6.21}$$

which has minimal variance on scales where $\xi(r) << 1$.

It has been shown that the optimal weighting of galaxies, for a precise measurement of the BAO peak, is to assign a weight to each galaxy Reid et al. [40]:

$$w_{gal} = (w_{cp} + w_{noz} - 1) \times w_{star} \times w_{see} \times w_{FKP} \,, \tag{6.22}$$

Here, the close-pair weight, w_{cp}, accounts for the fact that, due to fiber coating, one cannot assign optical fibers on the same plate to two targets closer than $62''$. The w_{noz} weight accounts for targets for which the pipeline failed to measure the redshift. The w_{star} and w_{see} weights correct for the dependance of the observed galaxy number density with the stellar density and with seeing, respectively. Finally, we use the FKP weight, w_{FKP}, Feldman et al. [41] in order to reduce the variance of the two-point correlation function estimator. Since we have introduced the random pairs in our analysis we have no longer access to the counts-in-spheres, but to the normalised counts-in-spheres

$$\mathcal{N}(< r) = N_{data}(< r)/N_{randoms}(< r) \,. \tag{6.23}$$

However, we can directly compute the normalised counts-in-spheres from the correlation function:

$$\widehat{\mathcal{N}}(< r) = 1 + \frac{3}{r^3} \int_0^r \hat{\xi}_{ls}(s) s^2 ds \,. \tag{6.24}$$

DRAFT

It has been shown that this estimator is expected to be the most optimal by Ntelis et al. [29]. Applying the previous result to equation Eq. 6.11 our estimator for the fractal correlation dimension is given by:

$$\widehat{D_2} = 3 + \frac{d\ln}{d\ln r} \left[1 + \frac{3}{r^3} \int_0^r \hat{\xi}_{ls}(s) s^2 ds \right] \,. \tag{6.25}$$

Throughout this document, we drop the hats for sake of simplicity. Estimators of the theoretical predictions of these observables are publicly available at COSMOlogical Python Initial Toolkit (COSMOPIT).

7 Limits on information mining

What can we extract from these observables? What information about the universe is accessible to us, directly or indirectly? Are there fundamental limits that underlie to our knowledge? Generally, the intrinsic limits-to the information we can have access to-are due to the finite speed of propagation of the information asserted by special relativity. The information that we extract comes from the photons arriving to our devices. This phenomenon gives also rise to the cosmic bias. Finally, an additional complications arise from the peculiar motion of the celestial objects on the sky. In this section, we are going to describe the effect of *causal diagrams*, the *cosmic bias* and the *redshift space distortions*. This section is based on Leclercq et al. [42] and Ntelis [10].

7.1 Causal diagrams

According to special relativity, there is a causal structure of the universe that is relevant for cosmology. In fact, it is only possible to observe part of the universe at a given time. This fact limits the information available for making statistical statements about scales comparable to the entire observable Universe. Since only a single realization from the ensemble of universes is accessible to us, statements about the largest scales are subject to uncertainty, usually referred to as cosmic variance. Causal diagrams are a convenient tool to visualize the information accessible directly or indirectly. They depict relativistic light cones whose surfaces describe the temporal evolution of light rays in space-time. These diagrams include both a future part (everything that you can possibly influence) and a past part (everything that can possibly have influenced you). On causal diagrams, your world line, i.e. your trajectory in space-time, is essentially a straight line orthogonal to the spatial space (the t-axis), for a stationary observer. As usual, for graphical convenience, we will suppress the three spatial dimension and represent the four-dimensional space-time in 2+1 dimensions. In addition, we will use comoving coordinates to factor out the expansion of the Universe, so that light-rays travel on diagonal lines. To examine further the causal structure of our universe, we will successively consider three categories: the information we can access now, directly; the information we could access directly, in a Universe's lifetime and the information we can access indirectly.

Information accessible directly, now. - Causality allows direct access, now, to:

- the surface of your past light cone (a 3D volume): all the photons that reach you now (e.g. photons from distant galaxies or from the CMB),
- the interior of your past light cone (a 4D volume): all events that could possibly influence you via a slower-than-light signal (this includes the gravitational field from massive objects or cosmic particles that you receive from space).

DRAFT

Fig. 5 illustrates the 3D lightcone and the information you can have access to directly. Your "CMB circle? (the last scattering sphere in 3D) is the intersection of a plane (3D Volume in reality) − corresponding to the time of last scattering sphere $t = t_{lss}$, i.e. the time when the CMB was emitted − and your past light cone.

Information accesible directely, over time - With the passing of time, progressively we receive more information through light. If one takes a telescope and gaze the galaxies at a fixed time, $t = t_1$, he will obtain information from the 4D lightcone of this particular time. At each moment of his world line there is a 4D lightcone. The more time he observes, the more 4D lightcones he observes and the more information he obtains. Note that for a finite age of the universe, at any given time, there are regions of the $t = 0$ 3D plane that we have not yet observed. Take the CMB for instance. The CMB we have access to changes with time, because the intersection of the 3D plane corresponding to the time of last scattering sphere $t = t_{lss}$, and the 3D lightcone changes when we consider a different lightcone. This means that in principle, waiting (for a long time!) allows access to a thick ring in the last scattering plane, i.e. the CMB map turns into a 3D CMB map.

Information accesible inderectly - If we want information that it is not directly accessible to us and we do not want to wait there is another option. The naive way is to proceed also indirectely. Knowing the laws of physics we can infer the behaviour of the universe in different regions of the spacetime continuum. Essentially, we can evolve observations forward and backward in the 4D volume and predict events in the interior (evolution backwards) or exterior (evolution forward) of our 3D lightcone. However such studies are model dependent

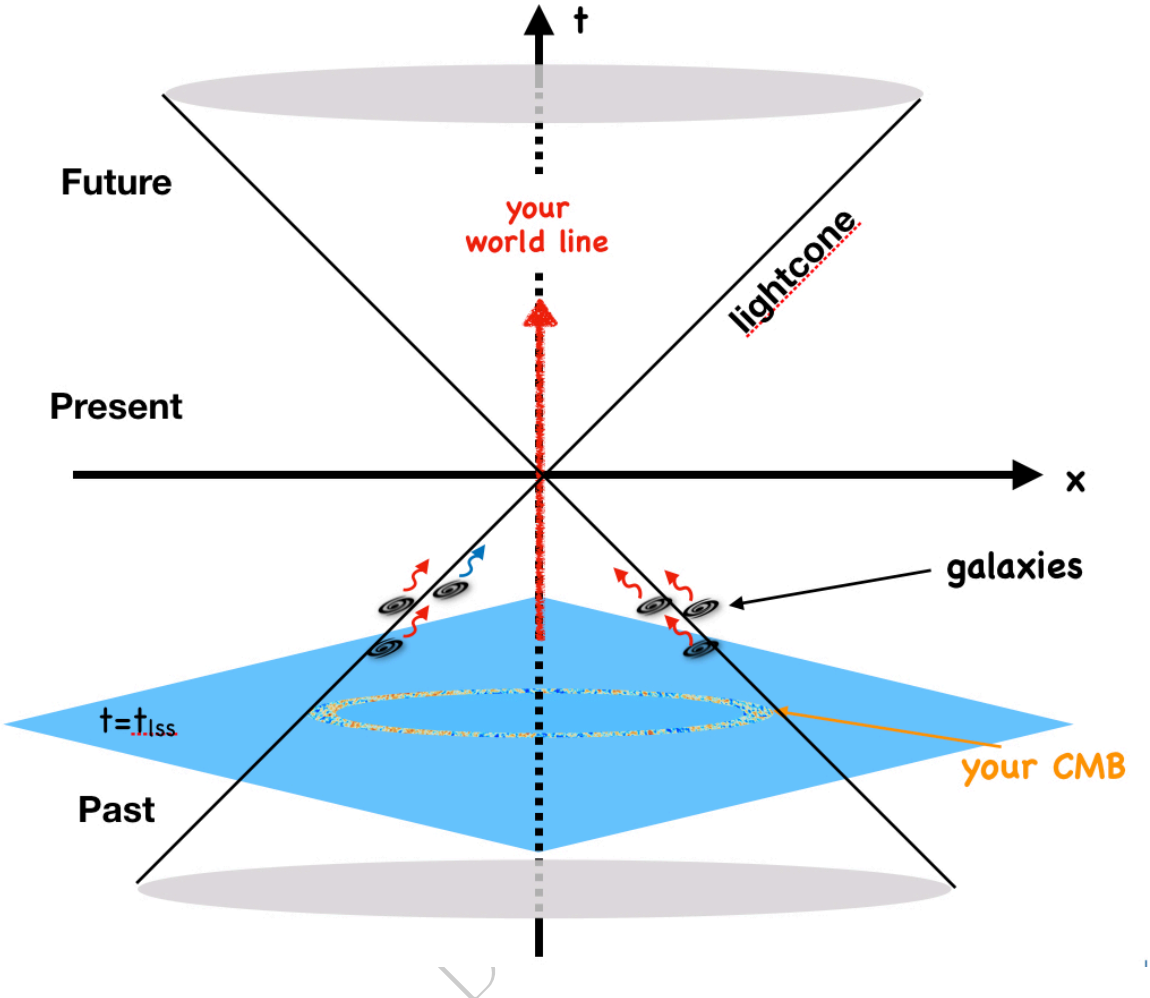

Figure 5. Causal diagram of the spacetime continum. The information that we have access at each time is only available in the interior of the 4D volume inside the 3D lightcone. The information that we observe each moment through light propagation is only available in the 3D lighcone, depicted here in 2 dimensional space. [Image inspired by Leclercq et al. [42]]

and it is difficult to test against the actual observational data. For a further discussion the interested reader is encouraged to read the review from Leclercq et al. [42] where they present some interesting ideas on how to get those observations.

7.2 Cosmic Bias

The ΛCDM model describes a universe filled with Cold Dark Matter and Dark Energy. Our current understanding suggests that the Dark Energy is responsible about the observed accelerating expansion nature of our universe. Cold Dark Matter is already observed as the missing mass of the galaxies when we study their rotational curves, as was explained in section 2. Therefore, we understand that the universe has a total matter distribution spread all over the spacetime continuum.

However what we are only able to observed with our telescopes is the light coming from galaxies, as explained in section 7.1. Thus what we are only have access to is the galaxy distribution, and therefore we can exctract the clustering statistis of galaxies, and not the

total matter of the universe. It turns out, when we compare the clustering statistic of our tracers and compare it with the theoretical prediction of the clustering statistic of the total matter of the universe, we end up having a biased relation between the two. The common statistic that defines this biased relation is usually given through the two point correlation function:

$$\xi_{tracer}(r) = b^2 \xi_{matter}(r) \tag{7.1}$$

where $\xi_{tracer}(r)$ is the two point correlation function of our tracer, usually the galaxy distribution, $\xi_{matter}(r)$ is the two point correlation function of the total matter of the universe and b is the *cosmic bias* or *bias*. There is an active field of research of exploring models of different kinds of biases, such as the local bias, or scale dependent biases[37].

7.3 Redshift Space Distortions

When we make the 3D map of the universe using the galaxy surveys, we are interested on the comoving angle positions (2D) of the galaxies in the sky and on their comoving radial position in the sky. However what we measure is their velocities through their luminosities, through the flux of photons in our devices. In particular, what we want to measure is the redshift, z, that was introduced in section 2. Let us explain what we measure in order to obtain the quantity of the redshift.

Firstly, to obtain the redshift, we measure the peaks of the luminosities of the individual galaxies as a function of wavelength or frequency of the incident photons in our devices. We identify the wavelength difference between those peaks of luminosities and we compare them with the wavelength differences of peaks of luminosities on well studied chemical elements in our laboratories, with the most notable one the Hydrogen! Therefore the observed redshift is basically the dopler effect of the galaxies in respect us. We called it redshift (because the most galaxies are going away from us and therefore their color shifts to redder colours.) Due to the expansion of the universe, there are additional swifts, on the redshifts that we observe for those galaxies. Therefore the observed redshift is given by:

DRAFT

$$z_{obs} = z_{pec} + z_{exp} \ , \tag{7.2}$$

where z_{obs} is the observed redshift, z_{pec} is the redshift due to the peculiar motion of the galaxy, and z_{exp} is the redshift due to the expansion of the universe. Note that always the measurements of redshifts are performed in the radial direction and in respect of us, and us we mean the local enviroment of the earth and our solar systems which we are currently able deploy our telescopes for those observations. Therefore this phenomenon produces a distortion on the actual expansion redshift that we would like to observe.

In terms of positions, the radial distance between us and a galaxy differs from the real-space due to the peculiar motion of the galaxies and the expansion. Therefore the actual radial position we observed is given by:

$$\vec{s}(r) = \vec{r} - v_r(r)\frac{\vec{r}}{r} \tag{7.3}$$

where $\vec{s}$ is the redshift radial comoving distance, $\vec{r}$ is the real comoving distance and v_r is the peculiar velocity in the radial direction.

Usually we are interested in two key regimes for redshift space distortions on large scale structure clustering studies; The linear regime, at scales $r \simeq [20 - 40]\ h^{-1}$ Mpc, and the non-linear regime, $r \lesssim 10\ h^{-1}$ Mpc. The linear regime is described by the Kaiser model[31],

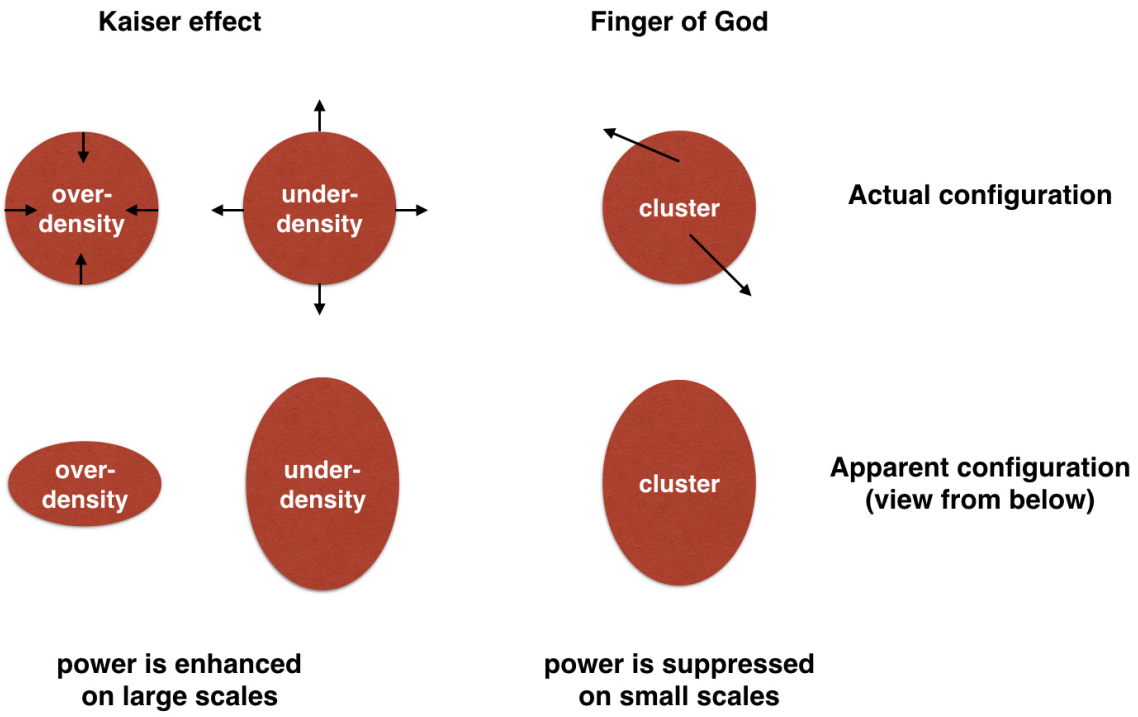

Figure 6. Illustration of Redshift Space distortions. Left: Kaiser effect. Right: Finger of God effect. Top: actual configuration. Bottom: apparent configuration as observed by a distant observer. [See text for details]

DRAFT

while for the non-linear regime there are several descriptions and it is an active research field in cosmology.

The kaiser model models the the velocity field of galaxies due to the fact that the galaxies are within a large over-density. In this over-density the gravitational pull is so large that galaxies are drifting within the center as shown in upper left part of Fig. 6. The perpendicular to the line of sight remains the same for the distant observer. The parallel to the line of sight produces a distortion due to this phenomenon. Therefore the image looks squeezed as shown by the bottom left part of Fig. 6. Kaiser [31] has shown that this phenomenon can be described by a model on the power spectrum of each tracer (galaxy, quasar or else) as:

$$P_{tracer}(k,\mu) = b^2(1+\beta\mu^2)P_{matter}(k) \tag{7.4}$$

with

$$\beta = f/b, \qquad f = \frac{d\ \ln D[a(t)]}{d\ln a(t)} \simeq \Omega_m^{\gamma}(z)\ , \tag{7.5}$$

where the last equality is a valid parametrization of the rate of growth of structures, f and $\gamma = 0.55$ for the unmodified General Relativity as shown by Linder and Cahn [43]. Notice that:

$$\Omega_m^{\gamma}(z) = \left[\frac{\Omega_m(1+z)^3}{(\Omega_m(1+z)^3 + 1 - \Omega_m}\right]^{\gamma} \tag{7.6}$$

for a flat ΛCDM model.

The non-linear regime is described empirically by the "Finger of God" effect which models by the dispersion of the peculiar velocity field of the galaxies at the cluster level, few kpc. In the litarature there are several models [44], so someone can take an compare these model according to their data. The Finger of God effect supresses the power spectrum at small scales, i.e. at large k modes due to the peculiar motions of the galaxies. Therefore the *Damping models* are summarised by:

$$D_G(k,\mu;\sigma_p,H_0) = exp\left[-\left(\frac{k\mu\sigma_p}{H_0}\right)^2\right] \text{, Gaussian} \tag{7.7}$$

and the

$$D_L(k,\mu;\sigma_p,H_0) = \frac{1}{1+\frac{1}{2}\left(\frac{k\mu\sigma_p}{H_0}\right)^2} \text{, Lorentzian} \tag{7.8}$$

These dumbs the theoretical model at smaller scales $r < 10\ h^{-1}$Mpc according to:

$$P_{tracer,final}(k,\mu;b,\sigma_p;H_0) = D_X(k,\mu;\sigma_p,H_0)P_{tracer}(k,\mu) \tag{7.9}$$

where D_X is either the *Lorentzian* model or the *Gaussian* model for damping.

8 Large scale structure surveys, status

Large scale structure surveys are usually devided into two big categories. The ones which focus on studying the primordial universe and the ones that focus on studying the late time universe. Often we use the their results to extract the physical information that explains both regimes of observations. The model that we are trying to describes is the ΛCDM model that explains the universe as a whole. The primordial universe observations are focus on studying the primordial temperature fluctuations. They study the Cosmic Microwave Background and there are several interesting physics going on.

DRAFT

Here we are going to the second part of observations, the ones that study the late time universe. At that time there are several models about the structure formation, content of the universe as long as alternative models of Gravity and many more. These observations are achieved by the so called, "large scale structure surveys". To perform such observations, we use telescopes that basically map the two dimensional position on the sky of several objects, as long as the luminosities of those objects. From the luminocities we can deduce velocities and radial distances and we can do a lot of interesting science with them! The basic observable that we use is the three Dimensional comoving density field and many of the reconstructions of that we tryied to summarise in section 6. Let's see what are the main instruments that are useful to extract the above information. The main characteristic that we are interested in cosmology, and large scale structure physics, is the *redshift* of each object that we described in section 2, therefore these surveys are called *Redshift Surveys* or *Galaxy Surveys*, since the main targets are galaxies.

The *Redshift Surveys* are usually devided into ground based and satellites. However, the satellite redshift surveys is a relative new theme of instruments that we are currently investigating, with the ongoing mission named, "Euclid Mission". The advantage of ground based observations is that we can easily modify the instrument, as well as the ongoing process of constructing it and testing it against observations is faster. However this allow for a very vast spectrum of different kind of instruments. However, they are subtle to atmospheric noise.

Therefore, for the first time, Euclid Mission is prepered to study the large scale structure in the late universe from space. The other important devision of instruments is that of Spectroscopy and Photometry that we will describe in the next two paragraphs.

In Photometry, instruments are build in such a way to take fast images of the sky. Photons transverse coloured filters and map the spectrum (flux as a function of wavelength) of the observed targets. The advantages are that they can perform fast imaging, with a good Signal to noise ratio (SNR), since one simple detector pipeline is used. Therefore they are able to perform massive data collection. However, their disadvantage is that they have a very small resolution on the wavelength.

In Spectroscopic instruments the photons transverse a sequences of dispersive materials so that we can map more precisely and accurately the spectrum of the targets. The advantages are that they perform an exquisite $\delta\lambda$ resolution. However, they have low SNR, since multiple detector elements are required of the photon pipeline. Furthermore, they perform slow imaging, since they require high exposure time. However, several robotic mechanisms are under development to ameliorate their speed and SNR.

Currently, the Sloan Digital Sky Survey (SDSS)[45] is observing the redshift region from $0 \leq z \leq 3.5$ targeting millions of galaxies and thousands of QSO with their corresponding Lyman-α forests. The main project is dedicated to cosmology and large scale structures. It has the name the extended Baryon Oscillation Spectroscopy Survey (eBOSS). The Dark Energy Spectroscopy Instrument (DESI)[46] is a dedicated project to study the large scale structure from the ground. We expect the first light in 2019. In 2019, we expect also the first light from the Large Synoptic Survey Telescope (LSST)[47] which is a sophisticate based photometric instrument with the a camera with the largest field of view for redshift surveys. Finally the Euclid Mission[48], is expected to give the first light in 2022, and is going to map the 3D comoving galaxy distribution in redshift $0.9 < z < 1.8$ and study the dark universe mainly from galaxy clustering and weak lensing. Currently, at CPPM[49], researcher are developing some characterisation of its NISP instrument[50]. All those instrument, target some overlapping regions to calibrate on one an another and explore as much as possible the vast universe!

DRAFT

9 Statistical inference

Information theory is a framework where the way of making decisions from a collection of data is studied [6]. One of the main things that we are interested from this framework is the statistical analysis or statistical inference. When discussing statistical data analysis, two different points of view are traditionally reviewed and opposed: the frequentist (see e.g. [51]) and the Bayesian approaches. It is commonly known that arguments for or against each of them are generally on the level of a philosophical or ideological position, at least among cosmologists today, 2018. Before criticizing this controversy, somewhat dated to the 20th century, and stating that more recent scientific work suppresses the need to appeal to such

[6] **Fun Fact**: In the framework of information theory the quote of Sokratis, "one think I know, that I known nothing", interpretes the equation $\mathcal{E}_X\left[I(X)\right] = \ln n$, where $I(x)$ is the self-information, which is the entropy contribution of an individual message, and $\mathcal{E}_X$ is the expected value for n messages. *Proof*: Let $\mathcal{X}$ be the set of all messages $\{x_1, \ldots x_n\}$ that an X random variable could be, and $p(x)$ is the probability of some $x \in \mathcal{X}$, then the entoropy, H, of X is defined as $\mathcal{E}_X\left[I(x)\right] = H(X) = -\sum_{x\in\mathcal{X}} p(x)\ln p(x)$. A property of entropy is that it is maximized when all the messages in the message space are equiprobable, $p(x) = 1/n$, i.e. the most unpredictable, in which case $\mathcal{E}_X\left[I(x)\right] = \ln n$.

arguments, we report the most common statements encountered. This section is based on Leclercq [52].

9.1 Bayesian vs Frequentist

Frequentist and Bayesian statistics differ in the epistemological interpretation of probability and their consequences for hypotheses testing and models comparison. Firstly, the methods differ on the understanding of the concept of the probability $P(A)$ of an event A. As a frequentist, one defines the probability $P(A)$ as the relative frequency with which the event A occurs in repeated experiments, i.e. the number of times the event occurs over the total number of trials, in the limit of a infinite series of equiprobable repetitions. This probability (definition) has several caveats. Besides being useless in real life (as it assumes an infinite repetition of experiments with nominally identical test conditions, requirement that is never met in most practical cases), it cannot handle unrepeatable situations, which have a particular importance in cosmology, as we have exactly one sample of the Universe. More importantly, this definition is surprisingly circular, in the sense that it assumes that repeated trials are equiprobable, despite that it is the very notion of probability that is being defined in the first place.

On the other hand, in Bayesian statistics, the probability $P(A)$ represents the degree of belief that any reasonable person (or machine) shall attribute to the occurrence of event A under consideration of all available information. This definition implies that in Bayesian theory, probabilities are used to quantify uncertainties independently of their origin, and therefore applies to any event. In other words, probabilities represent a state of knowledge in presence of partial information. This is the intuitive concept of probability as introduced by several authors such as Laplace, Bayes, Bernoulli, Metropolis, Jeffreys, etc.[53].

DRAFT

Translated to the measurement of a parameter in an experiment, the aforementioned definitions of probabilities yield differences in the questions addressed by frequentist and Bayesian statistical analyses. In the frequentist point of view, statements are structured as: "*the measured value x occurs with probability $P(x)$ if the measured quantity X has the true value X_T*". This means that the only questions that can be answered are of the form: "*given the true value X_T of the measured quantity X, what is the probability distribution of the measured values x?*". It also implies that statistical analyses are about building estimators, $\hat{X}$, of the truth, X_T.

In contrast, Bayesian statistics allows statements of the form: "*given the measured value x, the measured quantity X has the true value X_T with probability P*". Therefore, one can also answer the question: "*given the observed measured value x, what is the probability that the true value of X is X_T?*", which arguably is the only natural thing to demand from data analysis. For this reason, Bayesian statistics offers a principled approach to the question underlying every measurement problem, of how to infer the true value of the measured quantity given all available information, including observations. In summary, in the context of parameter determination, the fundamental difference between the two approaches is that frequentist statistics assumes the measurement to be uncertain and the measured quantity known, while Bayesian statistics assumes the observation to be known and the measured quantity uncertain. Similar considerations can be formulated regarding the problems of hypothesis testing and model comparison.

9.2 Bayesian Framework

The "plausible reasoning" can be formulated mathematically by introducing the concept of conditional probability $P(A|B)$, which describes the probability that the event A will occur given the information B which is given on the right side of the vertical conditioning bar "|". To conditional probabilities applies the following famous identity, which allows to go from forward modelling to the inverse problem, by noting that if one knows how x arises from y, then one can use x to constrain y:

$$P(y|x)P(x) = P(x|y)P(y) = P(x, y) \tag{9.1}$$

This observation forms the basis of Bayesian statistics.

Therefore, Bayesian analysis is a general method for updating the probability estimate for a theory in light of new data. It is based on Bayes' theorem,

$$P(\theta|d) = \frac{P(d|\theta)P(\theta)}{P(\theta)} , \tag{9.2}$$

where θ represents the set of the parameter space of a particular model of a particular theory and d represents the data or evidence (before the data are known). The above formula is interpreted as follows:

- $P(d|\theta)$ is the probability of the data before they are known, given the theory. It is usually called the *likelihood*.
- $P(\theta)$ is the probability of theory in the absence of data. It is called the prior probability distribution function or simply the *prior*.
- $P(\theta|d)$ is the probability of the theory, after the data are known. It is called the posterior probability distribution function or simply the *posterior*.
- $P(d)$ is the probability of the data before they are known, without any assumption about the theory. It is called the *evidence*.

DRAFT

One can think that the probability distribution function (pdf) for an uncertain parameter can be thought as a "belief distribution function", quantifying the degree of truth that one attributes to the possible values for some parameter. Certainty can be represented by a Dirac distribution, e.g. if the data determine the parameters completely.

In summary, the inputs of a Bayesian analysis are two:

- the data: include for example, the galaxy angle position in the sky, galaxy redshift, photometric redshift pdfs, the temperature in pixels of a CMB map, etc. Details of the survey specifications have also to be accounted for at this point: noise, mask, survey geometry, selection effects, biases, etc.
- the prior: it includes modelling assumptions, both theoretical and experimental. Specifying a prior is a systematic way of quantifying what one assumes true about a theory before looking at the data.

While the output of a Bayesian analysis is the posterior density function. The prior choice is a key ingredient of Bayesian statistics. It is sometimes considered problematic, since there is no unique prescription for selecting the prior. Here we can argue that prior specification is

not a limitation of Bayesian statistics and does not undermine objectivity. One can simply determine the prior knowledge that he consider and compare with studies that have used different prior. The discussion of selecting the appropriate prior is beyond the scope of these notes and the reader is redirected to Leclercq [52].

9.3 Statistical indeference problem

Data analysis problems can be typically classified as: parameter inference, model comparison, hypothesis testing. For example, cosmological questions of these three types, related to the large-scale structure, would be:

- What is the value of the dark energy density ratio, Ω_Λ?
- Is structure formation driven by general relativity or by modified gravity?
- Are large-scale structure observations consistent with the hypothesis of an inflationary scenario?

In this section, we describe the methodology for questions of the first two types. Hypothesis testing, i.e. inference within an uncertain model, in the absence of an explicit alternative, can be treated in a similar manner.

9.4 First level inference: parameter estimation

DRAFT

The general problem of parameter estimation can be stated as follows. Given a physical model M, a set of hypotheses is specified in the form of a vector of parameters, θ. Together with the model, priors for each parameter must be specified: $P(\theta|M)$. The next step is to construct the likelihood function for the measurement, with a probabilistic, generative model of the data: $P(d|\theta, M)$. The likelihood reflects how the data are obtained: for example, a measurement with Gaussian noise will be represented by a normal distribution.

Once the prior is specified and the data is incorporated in the likelihood function, one immediately obtains the posterior distribution for the model parameters, integrating all the information known to date, by using Bayes' theorem eq. 9.2:

$$P(\theta|d, M) = \frac{P(d|\theta, M)P(\theta|M)}{P(d)} \tag{9.3}$$

Note that the normalizing constant, namely *Bayesian evidence* is defined as:

$$P(d) = \int_M \int_\theta P(d|\theta, M) = \int_{M_1} \cdots \int_{M_n} \int_{\vec{\theta_1}} \cdots \int_{\vec{\theta_n}} p(d|M_1(\vec{\theta_1}) \dots M_n(\vec{\theta_n})) \tag{9.4}$$

where $\int_x f(x) = \int_X f(x)dx$ implies the usual Remannian integration. The Bayesian evidence is irrelevant for parameter inference (but fundamental for model comparison, see section 9.7). Usually, the set of parameters θ can be divided in some physically interesting quantities ϕ and a set of nuisance parameters n. The posterior obtained by eq. 9.3 is the joint posterior for $\theta = (\phi, n)$. The marginal posterior for the parameters of interest is written as (marginalizing over the nuisance parameters):

$$P(\phi|d, M) \propto \int P(d|\phi, n, M)P(\phi, n|M)dn \; . \tag{9.5}$$

This pdf is the final inference on ϕ from the joint posterior. The following step, to apprehend and exploit this information, is to explore the posterior. One usually uses Monte Carlo Markov Chains (MCMC) to explore the posterior and then he can represent the result in the form of the posterior probability in the form of 1D plots or 2D contour plots. A more detail discussion on the estimation of the posterior using MCMC is given in section 9.9.

9.5 Example 1

Let's take an example. Given the likelihood:

$$P(d|\alpha,\beta) = \left[\frac{\alpha-1}{0.3}\right]^2 + \left[\frac{\beta-1}{0.4}\right]^2 \tag{9.6}$$

Notice that the data are equal only 0. In the figure 7, the schematic representation of the posterior of the above model is given in 3 panels. This schematic representation is usually referred to as *corner plot*. In the top panel the posterior probability distribution function is plotted for α parameter (1D plot). On the right corner of this panel the mean and 1σ standard deviation is plotted. Notice that the posterity distribution is normalised so that its maximum is 1. The same for the posterior of the β parameter in the right corner on this figure. In the left bottom corner one can see the 2D plot of the joint probability density distribution of the model. The hight of the probability distribution is denoted usually with the density of the chains.

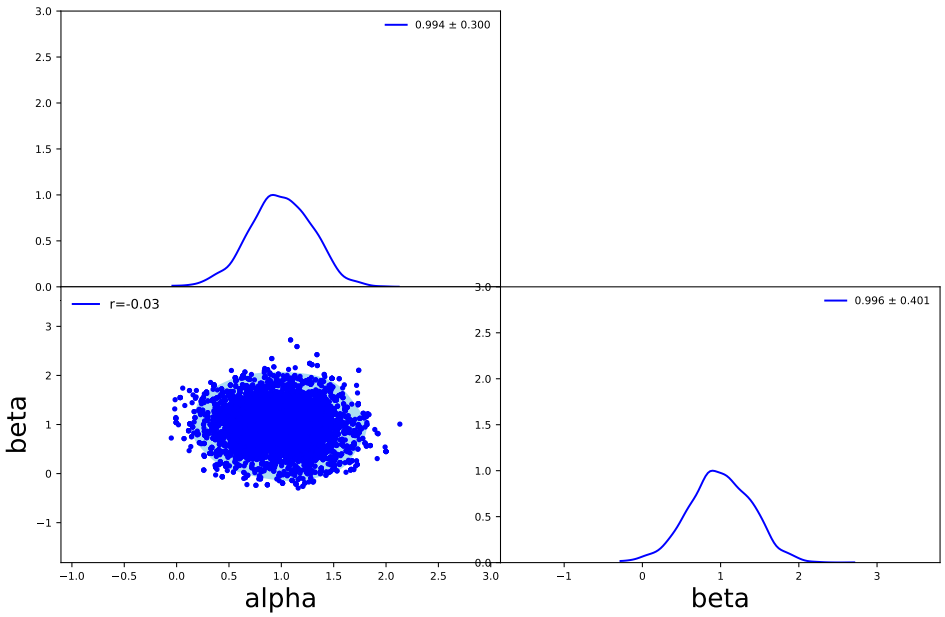

Figure 7. Representation of the posterior distribution of an 1D and 2D parameter model. [See text for details]

9.6 Example 2

Now lets take a more realistic case by taking another example using data. Since we do not possess any data in our disposal, lets construct some fake ones, X_D, Y_D, σ_D. Now we can construct our simple model:

$$y_m = ax + b \tag{9.7}$$

lets call it *linear model*. Now we can construct our Prior:

$$\ln P(a,b) = \left[\frac{a-1}{0.05}\right]^2 + \left[\frac{b-0.05}{0.05}\right]^2 := \ln P_{Prior} \tag{9.8}$$

Now we can construct the likelihood:

$$\ln P(d|a,b) = \sum_{i=0}^{N-1} \left[\frac{y_D(x_i) - y_m(x_i|a,b)}{\sigma_{y_D}}\right]^2 := \ln P_{Data} \tag{9.9}$$

If we want we can combine the Prior information with the information coming from the data using the following way:

$$\ln P_{Data+Prior} = \ln P_{Data} + \ln P_{Prior} \tag{9.10}$$

which can be writtien in the explicit form :

$$\ln P_{DP}(d|a,b) = \ln P(d|a,b) + \ln P(a,b) \tag{9.11}$$

DRAFT

Now we can maximize the logarithm of the likelihood to find the best value for α, β.

By using an MCMC algorithm [54] we can easily maximize equations 9.8, 9.9 and 9.11. The results of the fitted models are shown in figure 8 where the reduced χ^2 is given.

NOTE: The reduced χ^2 is the χ^2 devided by the degrees of freedom. The degrees of freedom are the number of bins of the data minus the number of free parameters of the model. What do we observe? Notice that the Prior information was a better fit to the data. Then using only the Data the χ^2 is small. and deviates a lot from the degrees of freedom. By using both data and the prior information you can see that the χ^2 is increasing which is an indication of a better fitting. So by using the Prior knowledge we enhance the fitting of our model.

In figure 9 we can observe the resulting contours for the different method of estimations using only the Prior (blue color), using only the Data (green color) and using both Prior and Data (red color). Notice that using the prior the precision on the estimation of the parameters is enhanced, i.e. the errors of the parameters are getting smaller. Therefore we have more precise measurement.

9.7 Second level inference: model comparison

Contrary to the frequentist's approach where the model comparison is simply inferred by the comparison the χ^2, second level Bayesian inference always requires an alternative explanation for comparison (finding that the data are unlikely within a theory does not mean that the theory itself is improbable, unless compared with an alternative).

The evaluation of model M's performance given the data is quantified by $P(M|d)$. Using Bayes' theorem to invert the order of conditioning, we see that it is proportional to

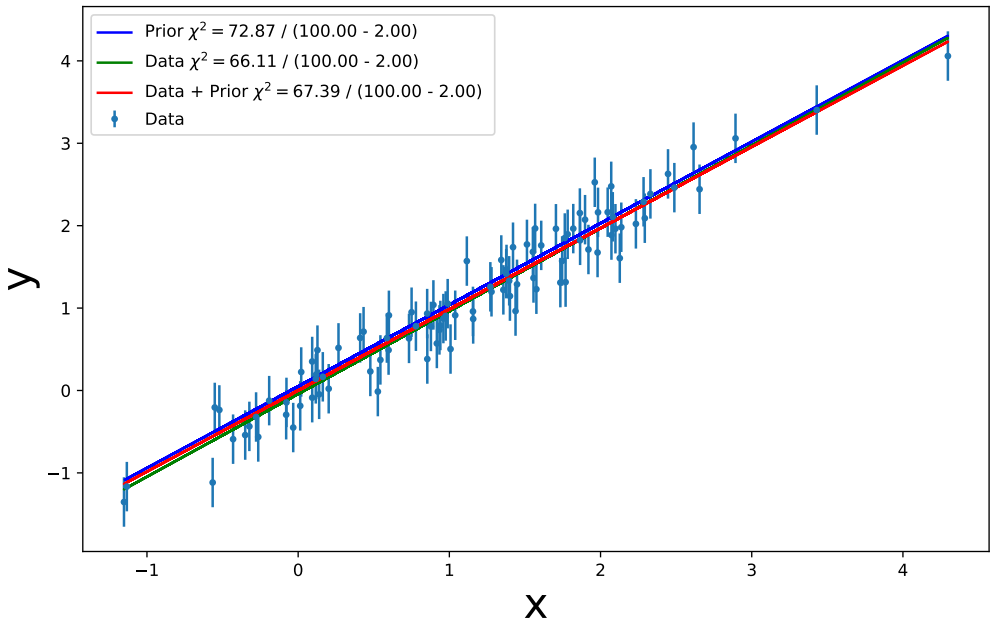

Figure 8. The linear fitting Using the MCMC method. Light blue are the data. Dark Blue is the best fit model using only the prior. With Green is the best fit model using only the data. With red is the best fit model using both the prior and the data.

DRAFT

the product of the prior probability for the model itself, $P(M)$, and the Bayesian evidence already encountered in the first level inference, $P(d|M)$:

$$P(M|d) \propto P(M)P(d|M) \; . \tag{9.12}$$

(It is implied that those quantities have integrated out the dependence on the parameters of the model, $\vec{\theta}$). Usually, prior probabilities for the models are taken as all equal to $1/N_m$, where N_m are the different models (this choice is said to be non-committal). When comparing two competing models denoted by M_0 and M_1, one is interested in the ratio of the posterior probabilities, given by:

$$P_{01} := \frac{P(M_0|d)}{P(M_1|d)} = \frac{P(M_0)P(d|M_0)}{P(M_1)P(d|M_1)} \tag{9.13}$$

With non-committal priors on the models, $P(M_0) = P(M_1)$, the ratio simplifies to the ratio of evidences, called the *Bayes factor*:

$$\mathcal{B}_{01} := \frac{P(d|M_0)}{P(d|M_1)} \tag{9.14}$$

The Bayes factor is the relevant quantity to update our state of belief in two competing models in light of the data, regardless of the relative prior probabilities we assign to them: a value of $\mathcal{B}_{01}$ greater than one, i.e. $\mathcal{B}_{01} > 1$, means that the data support model M_0 over model M_1. Note that, generally, the Baye's factor is very different from the ratio of likelihoods: a more complicated model will always yield higher likelihood values, whereas the evidence will

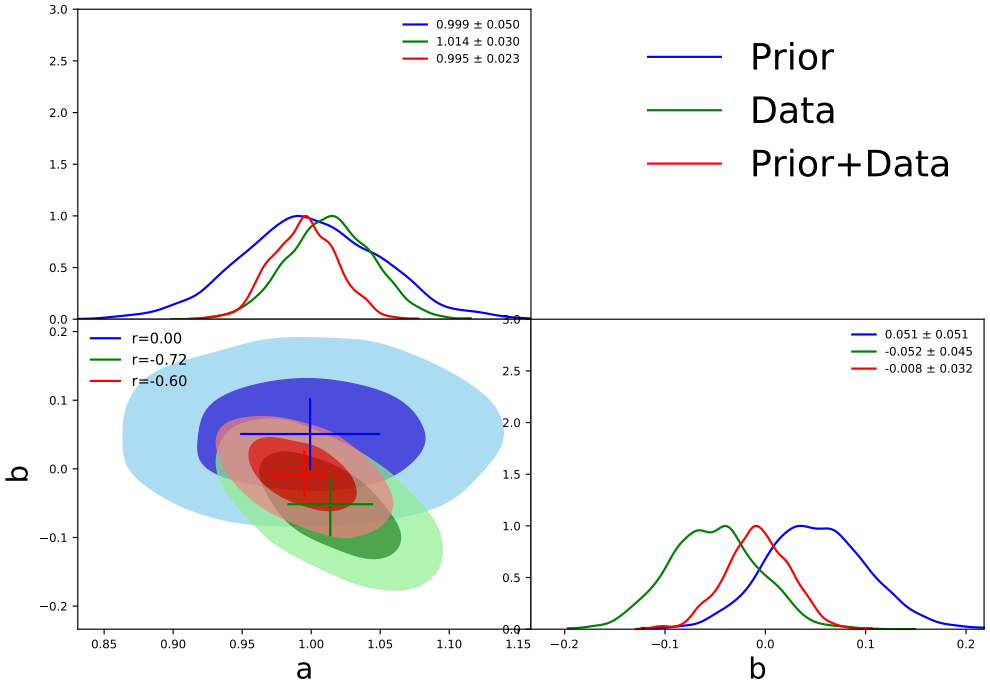

Figure 9. Is the resulting contour plot of figure 8. Dark contours represent the 1σ region (68% C.L.) the light contours represent the 2σ region (95% C.L.).

DRAFT

favor a simpler model if the fit is nearly as good, through the smaller prior volume, (*Occam's Razor*).

In practice one uses the so called Akaike Information Criterion corrected (AICc) [55] and Bayesian Information Criterion (BIC)[56] tests for model comparison. For N_D the number of data, N_θ number of parameters and the resulting χ^2 from the fitted model we have the following. The AICc is defined as:

$$AICc = \chi^2 + 2N_D + \frac{2N_\theta^2 + 2N_\theta}{N_D - N_\theta - 1.} \tag{9.15}$$

While the BIC is defined as:

$$BIC = \chi^2 + 2N_\theta \ln(N_D) \tag{9.16}$$

The model with the smallest statistic, AICc or BIC, is the preferred description of the data. Notice, how we penalise the models that have large number of parameters N_θ (*Occam's razor*). Therefore when $AICc(M1) < AICc(M2)$, we keep the model M1 and we discard the model M2.

9.8 Example 3

Now going back to our example, section 9.5, we can fit our data with a more sophisticated model that can follow the complexity of the data. In this case we can choose the *spline model*[57]. This model is a piece wise polynomial with gaussian components. The can choose

a high number of nodes for this model $N_\theta^{spl} = 27$. Therefore we fit the spline model to the data as well and we present the resulting metric to figure 10.

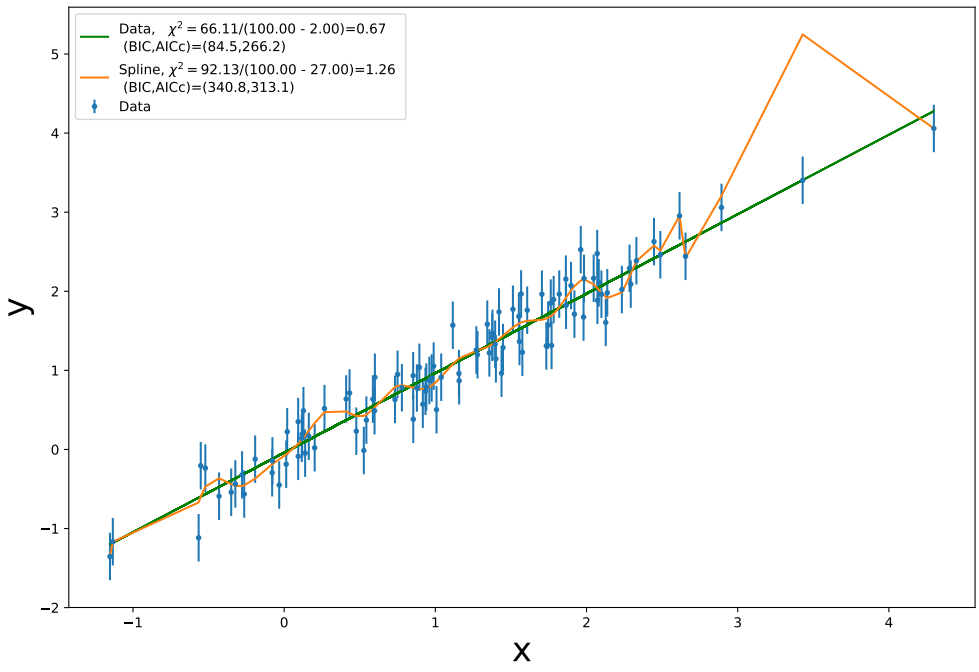

Figure 10. The data are denoted with light blue. The linear model (green line) and the spline model (green line) are compared against the data. [See text for details].

As you can observe, the resulting model follows better the data rather than the linear model, given by equation 9.7, with $N_\theta^{linear} = 2$ parameters. Even though the resulting χ^2_{spline} is closer to 1 than the χ^2_{linear} the AICc and BIC criteria shows otherwise. The $AICc(spline)$, $BIC(spline)$ are larger than $AICc(linear)$, $BIC(linear)$. Therefore the *linear model* is preferred to the more sophisticated one, i.e. *spline model*.

9.9 MCMC parameter exploration

Usually, the list of parameters is long, and thus multi-parameter likelihood calculations would be computationally expensive using grid-based techniques. Consequently, fast methods to explore parameter spaces are popular, particularly the Markov-Chain Monte-Carlo (MCMC) technique, which is commonly used for such analyses. While there is publicly available code to calculate cosmological model constraints [23, 24], the basic method is extremely simple and relatively straightforward to code.

9.10 Basic Algorithm

The MCMC method provides a way to generate a random sequence of parameter values whose distribution matches the posterior probability distribution. These sequences of parameter, or *chains* are commonly generated by an algorithm called the Metropolis-Hasting algorithms [58]. The algorithm is as folows: given a value at position θ, a candidate point θ_p is chosen at

random from a proposal distribution $f(\theta_p|\theta)$ - usually by means of a random number generator tuned to this proposal distribution. Then the algorithm has to decide if it will move to the candidate point. This transition is accepted if the new position has a higher likelihood. If the new position θ_p is less likely than θ, then we must draw another random variable, this time with uniform density between 0 and 1. The θ_p is accepted, and the chain moves to point θ_p, if the random variable is less than the ratio of the likelihood of θ_p and the likelihood of θ. Otherwise the chain "stays" at θ, giving this point extra weight within the sequence. In the limit of an infinite number of steps, the chains will reach a converged distribution where the distribution of chain links are representative of the hyper-surface of the likelihood, given any symmetric proposal distribution $f(\theta_p|\theta) = f(\theta|\theta_p)$.

It is common to implement dynamic optimisation of the sampling of the likelihood surface[59], performed in a period of burn-in at the start of the process. The convergence is always an issue. How do we know when we have sufficiently long chains that we have adequately sampled the posterior probability. A number of tests are available [60]. In the next section 9.11, we describe a test that allows to find the convergence of an MCMC by obtaining the results from different chains started at widely separated locations in parameter space.

9.11 Tests of MCMC convergence

In order to see if an MCMC parameter estimation algorithm has been converged one usually uses several methods. These methods can been classified to the "visual" inspection and the calculation of different statistics. Visual inspection includes that the walkes of an MCMC oscillate around a common, mean value. The different statistics that are used are the Gelman-Rubin Test [60], the Heidelberger-Welch test[61], the Goodman-Weare test [62]. Here, we are going to describe the Gelman-Rubyn test. For an overview of the rest of the tests, see of the Appendix 6.B of [63].

DRAFT

The statistic[60] which is called the Gelman-Rubyn statistic is denoted usually as R. How one computes that? One need to follow the forthcoming steps:

- Draw m independent MCMC realisation of a set parameters p_i of a model after n steps of each chain.
- The new parameters now can be denoted as θ_{ij} where $i = [0, ..., n]$ and $j = [0, ..., m]$
- Then computes the quantities the $R(\theta)$ for each parameter, θ, as follows:

The mean $\bar{\theta_j}$ and standard deviation σ_j of each parameter, for each chain, j:

$$\bar{\theta_j} = \frac{1}{n}\sum_{i=1}^{n}\theta_{ij}, \ ; \sigma_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\theta_{ij} - \bar{\theta_j})^2 \tag{9.17}$$

Then one computes the derived quantities: The mean of the mean

$$\bar{\bar{\theta}} = \frac{1}{m}\sum_{j=1}^{m}\bar{\theta_j} \tag{9.18}$$

The mean of the variances:

$$W = \frac{1}{m}\sum_{j=1}^{m}\sigma_j^2 \tag{9.19}$$

The deviations of the mean of the chains from the mean of the of the means:

$$B = \frac{n}{m-n}\sum_{j=1}^{m}(\bar{\theta_j} - \bar{\bar{\theta}})^2 \tag{9.20}$$

Then the quantity:

$$V(\theta) = (1 - 1/n)W + B/n \tag{9.21}$$

Finally the Gelman-Rubyn statistic is given by:

$$R(\theta) = \sqrt{\frac{V(\theta)}{W}} \tag{9.22}$$

Compute this statistic with starting from a different iteration of the chains. The Rule of thumb of convergence is $R - 1 < 0.03$ or for more demanding results $R - 1 < 0.01$.

10 Summary & Conclusion

We presented the basic concepts of physics on large scale structures. We gave a brief overview on the thermal history of our universe. We describe the theoretical framework, behind the magnificent scenery of the current observations, known as the standard ΛCDM model. We described the necessary tools for observations by introducing the statistical observables currently investigated. We discussed the current and future large scale structure surveys. Finally, we gave an overview on the statistical elements needed to study large scale structures, with a focus on Bayesian Analysis. Physics, mathematics, statistics and informatics are valuable skills! Believe it or not, keep on searching! ;-)

DRAFT

Acknowledgements

PN would like to thank Mikhail Stolpovskiy, Ranajoy Banerji, Cyrille Doux, as well as Jean-Christophe Hamilton and James Rich for fruitful and endless discussions on these subjects over his four years of research. PN would like to thank the organisers of *L'école d'éte France Excellence 2018* for such a pleasant environment that created for this event. PN is supported by a CNES Post Doctoral fellowship.

Bibliography

[1] Hubble, E. A relation between distance and radial velocity among extra-galactic nebulae. *Proceedings of the National Academy of Sciences*, 15:168–173, 1929.

[2] Planck Collaboration and Ade, P. A. R. and Aghanim, N. and Arnaud, M. and Ashdown, M. and Aumont, J. and Baccigalupi, C. and Banday, A. J. and Barreiro, R. B. and Bartlett, J. G. and et al., t. . P. 2016. âĂIJarXivâĂİ:1502.01589.

[3] Dodelson, S. *Modern cosmology*. Academic press, 2003.

[4] Einstein, A. Kosmologische und relativitatstheorie. *SPA der Wissenschaften*, 142, 1917.

[5] Alpher, R. A., H. Bethe, and G. Gamow. The origin of chemical elements. *Physical Review*, 73:803, 1948.

[6] Penzias, A. A. and R. W. Wilson. A measurement of excess antenna temperature at 4080 mc/s. *The Astrophysical Journal*, 142:419–421, 1965.

[7] Fixsen, D., E. Cheng, J. Gales, et al. The Cosmic Microwave Background Spectrum from the Full COBE FIRAS Data Set. *The Astrophysical Journal*, 473:576, 1996. astro-ph/9605054.

[8] Bennett, C., D. Larson, J. Weiland, et al. Nine-year wilkinson microwave anisotropy probe (wmap) observations: final maps and results. *The Astrophysical Journal Supplement Series*, 208:20, 2013.

[9] Gunn, J. E. and B. A. Peterson. On the density of neutral hydrogen in intergalactic space. *The Astrophysical Journal*, 142:1633–1641, 1965.

[10] Ntelis, P. *Probing Cosmology with the homogeneity scale of the Universe through large scale structure surveys*. Theses, Astroparticule and Cosmology Group, Physics Department, Paris Diderot University, 2017. URL https://hal.archives-ouvertes.fr/tel-01674537.

[11] Zwicky, F. On the masses of nebulae and of clusters of nebulae. *The Astrophysical Journal*, 86:217, 1937.

[12] Rubin, V. C. and W. K. Ford Jr. Rotation of the andromeda nebula from a spectroscopic survey of emission regions. *The Astrophysical Journal*, 159:379, 1970.

[13] Riess, A. G., A. V. Filippenko, P. Challis, et al. Observational evidence from supernovae for an accelerating universe and a cosmological constant. *The Astronomical Journal*, 116: 1009, 1998.

DRAFT

[14] Perlmutter, S., G. Aldering, M. della Valle, et al. Discovery of a supernova explosion at half the age of the universe. 1998. astro-ph/9712212.

[15] Schmidt, B. P., N. B. Suntzeff, M. M. Phillips, et al. The High-Z Supernova Search: Measuring Cosmic Deceleration and Global Curvature of the Universe Using Type IA Supernovae. 1998. astro-ph/9805200.

[16] Eisenstein, D. J., I. Zehavi, D. W. Hogg, et al. Detection of the baryon acoustic peak in the large-scale correlation function of sdss luminous red galaxies. *The Astrophysical Journal*, 633:560, 2005.

[17] Cole, S., W. J. Percival, J. A. Peacock, et al. The 2dF Galaxy Redshift Survey: power-spectrum analysis of the final data set and cosmological implications. 2005. astro-ph/0501174.

[18] McGaugh, S. S. The baryonic tully-fisher relation of gas-rich galaxies as a test of λcdm and mond. *The Astronomical Journal*, 143:40, 2012.

[19] Hamilton, J.-C. What have we learned from observational cosmology? *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 46:70–85, 2014.

[20] Baumann, D. Cosmology lectures by daniel bauman, 2018. http://www.damtp.cam.ac.uk/user/db275/Cosmology.pdf.

[21] Adam, R., P. Ade, N. Aghanim, et al. Planck 2015 results. i. overview of products and scientific results. 2015. arXiv:1502.01582.

[22] Eisenstein, D. J. and W. Hu. Power spectra for cold dark matter and its variants. *The Astrophysical Journal*, 511:5, 1999.

[23] Blas, D., J. Lesgourgues, and T. Tram. The cosmic linear anisotropy solving system (class). part ii: approximation schemes. *Journal of Cosmology and Astroparticle Physics*, 2011:034, 2011.

[24] Lewis, A. and A. Challinor. Camb: Code for anisotropies in the microwave background. *Astrophysics Source Code Library*, 2011.

[25] Percival, W. J. Large Scale Structure Observations. *ArXiv e-prints*, 2013. arXiv:astro-ph.CO/1312.5490.

[26] Labini, F. S., M. Montuori, and L. Pietronero. Comment on the paper by l. guzzo" is the universe homogeneous?". 1998. arXiv:astro-ph/9801151.

[27] Coleman, P. H. and L. Pietronero. The fractal structure of the universe. *Physics Reports*, 213:311–389, 1992.

[28] Scrimgeour, M. I., T. Davis, C. Blake, et al. The wigglez dark energy survey: the transition to large-scale cosmic homogeneity. *Monthly Notices of the Royal Astronomical Society*, 425:116–134, 2012. arXiv:1205.6812.

[29] Ntelis, P., J.-C. Hamilton, J.-M. Le Goff, et al. Exploring cosmic homogeneity with the BOSS DR12 galaxy sample. 2017. arXiv:1702.02159.

DRAFT

[30] Laurent, P., J.-M. Le Goff, E. Burtin, et al. A 14 h^{-3} Gpc^3 study of cosmic homogeneity using BOSS DR12 quasar sample. 2016. arXiv:1602.09010.

[31] Kaiser, N. Clustering in real space and in redshift space. *Monthly Notices of the Royal Astronomical Society*, 227:1–21, 1987.

[32] Hu, W. Angular trispectrum of the cosmic microwave background. *Physical Review D*, 64:083005, 2001.

[33] Castro, P. G. Bispectrum and the trispectrum of the ostriker-vishniac effect. *Physical Review D*, 67:123001, 2003.

[34] Carron, J. and I. Szapudi. What does the n-point function hierarchy of the cosmological matter density field really measure? *Monthly Notices of the Royal Astronomical Society*, 469:2855–2858, 2017.

[35] Karagiannis, D., A. Lazanu, M. Liguori, et al. Constraining primordial non-gaussianity with bispectrum and power spectum from upcoming optical and radio surveys. *Monthly Notices of the Royal Astronomical Society*, 2018.

[36] Amendola, L., S. Appleby, A. Avgoustidis, et al. Cosmology and fundamental physics with the euclid satellite. *Living Reviews in Relativity*, 21:2, 2018.

[37] Desjacques, V., D. Jeong, and F. Schmidt. Large-scale galaxy bias. *arXiv preprint arXiv:1611.09787*, 2016.

[38] Peebles, P. J. E. and M. G. Hauser. Statistical analysis of catalogs of extragalactic objects. iii. the shane-wirtanen and zwicky catalogs. *The Astrophysical Journal Supplement Series*, 28:19, 1974.

[39] Landy, S. D. and A. S. Szalay. Bias and variance of angular correlation functions. *The Astrophysical Journal*, 412:64–71, 1993.

[40] Reid, B., S. Ho, N. Padmanabhan, et al. Sdss-iii baryon oscillation spectroscopic survey data release 12: galaxy target selection and large-scale structure catalogues. *Monthly Notices of the Royal Astronomical Society*, 455:1553–1573, 2016.

[41] Feldman, H. A., N. Kaiser, and J. A. Peacock. Power spectrum analysis of three-dimensional redshift surveys. *arXiv preprint astro-ph/9304022*, 1993.

[42] Leclercq, F., A. Pisani, and B. Wandelt. Cosmology: From theory to data, from data to theory, 2014.

[43] Linder, E. V. and R. N. Cahn. Parameterized beyond-einstein growth. *Astroparticle Physics*, 28:481–488, 2007. arXiv:astro-ph/0701317v2.

[44] Ballinger, W., J. Peacock, and A. Heavens. Measuring the cosmological constant with redshift surveys. *arXiv preprint astro-ph/9605017*, 1996. arXiv:astro-ph/9605017v1.

[45] Blanton, M. R., M. A. Bershady, B. Abolfathi, et al. Sloan Digital Sky Survey IV: Mapping the Milky Way, Nearby Galaxies, and the Distant Universe. 2017. arXiv:1703.00052.

DRAFT

[46] Aghamousa, A., J. Aguilar, S. Ahlen, et al. The desi experiment part i: Science, targeting, and survey design. *arXiv preprint arXiv:1611.00036*, 2016.

[47] Collaboration, L. D. E. S. et al. Large synoptic survey telescope: dark energy science collaboration. *arXiv preprint arXiv:1211.0310*, 2012.

[48] Laureijs, R., J. Amiaux, S. Arduini, et al. Euclid definition study report. *arXiv preprint arXiv:1110.3193*, 2011.

[49] Renoir team of center of particle physics in marseille. https://www.cppm.in2p3.fr/renoir/.

[50] Nisp instrument in cppm. https://www.cppm.in2p3.fr/renoir/euclid_nisp.php.

[51] Kendall, M. G. and A. Stuart. *Vol. 3: Design and Analysis, and Time-series*. Charles Griffin Limited, 1968.

[52] Leclercq, F. *Bayesian large-scale structure inference and cosmic web analysis*. PhD thesis, U. Paris-Saclay, Orsay, 2015, arXiv:astro-ph.CO/1512.04985. URL https://inspirehep.net/record/1409879/files/arXiv:1512.04985.pdf.

[53] Jaynes, E. T. *Probability theory: the logic of science*. Cambridge university press, 2003.

[54] Pymc user's guide. https://pymc-devs.github.io/pymc/.

[55] Wikipedia contributors. Akaike information criterion — Wikipedia, the free encyclopedia, 2018. URL https://en.wikipedia.org/w/index.php?title=Akaike_information_criterion&oldid=839128392. [Online; accessed 1-May-2018].

[56] Wikipedia contributors. Bayesian information criterion — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Bayesian_information_criterion&oldid=833789064, 2018. [Online; accessed 1-May-2018].

[57] Wikipedia contributors. Spline (mathematics) — Wikipedia, the free encyclopedia, 2018. URL https://en.wikipedia.org/w/index.php?title=Spline_(mathematics)&oldid=837660470. [Online; accessed 1-May-2018].

[58] Wikipedia contributors. Metropolis?hastings algorithm — Wikipedia, the free encyclopedia, 2018. URL https://en.wikipedia.org/w/index.php?title=Metropolis%E2%80%93Hastings_algorithm&oldid=838047823. [Online; accessed 11-May-2018].

[59] Roberts, G. O., A. Gelman, W. R. Gilks, et al. Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, 7:110–120, 1997.

[60] Gelman, A. and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statist. Sci.*, 7:457–472, 1992.

[61] Heidelberger, P. and P. D. Welch. A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM*, 24:233–245, 1981.

[62] Goodman, J., J. Weare, et al. Ensemble samplers with affine invariance. *Communications in applied mathematics and computational science*, 5:65–80, 2010.

DRAFT

[63] Doux, C. *Combinations of cosmological probes: two applications with data from Planck and SDSS-III/BOSS*. PhD thesis, APC, Paris, 2017. http://inspirehep.net/record/1665198/files/fulltext.pdf.

DRAFT