**Lecture Notes for ICTP-Trieste/ICTP-SAIFR School on Cosmology**

# Structure formation in the universe

**Dragan Huterer, University of Michigan.**

**Introduction.** A cosmological theory (or, a model) predicts *statistical* properties of the distribution of objects on the sky. The theory does not predict the number density of galaxies in some direction in the sky, but rather it predicts the *distribution* of number-densities.

The distribution of objects is most generally described in terms of the infinite hierarchy of *correlation functions* of the positions of objects. The most famous and useful ones are:

1. The *number density* of objects, which is formally just the one-point correlation function. While counting objects on the sky seems an obvious way to compare theory to observations, it turns out it is very hard to theoretically predict the number of objects, except for the most massive ones – the *galaxy clusters*. Near the end of this chapter, we will cover the so-called *mass function*, $dn/d\ln M$, of the clusters.

2. The *two-point correlation function* (of galaxies, clusters etc) $\xi(r)$ – and its Fourer-space 'cousin', the *power spectrum* $P(k)$, turn out to be the most useful meeting point between observations and theory. As we will discuss at length in what follows, two-point function and/or the power spectrum are readily predicted and measured.

Higher-point correlation functions (esp. the three-point function, or its Fourier friend, the *bispectrum*) are also very useful and subject of a lot of research, but we do not cover them here, except to write a formal definition of the three-point function in real space in Eq. (11).

We start this chapter by introducing the fundamental variable in structure formation – the density fluctuation (or perturbation, or contrast), $\delta$.

**Density Perturbations.** We define an overdensity with the symbol $\delta$

$$\boxed{\delta(\mathbf{x},t) \equiv \frac{\rho(\mathbf{x},t)-\bar{\rho}}{\bar{\rho}}} \tag{1}$$

where $\rho(\mathbf{x},t)$ is the density (mass per volume) at any given location $\mathbf{x}$ and time $t$, and $\bar{\rho}$ is the mean density of all space. You can think of the density $\rho(\mathbf{x},t)$ as being defined in a small region of space – for example, within a small sphere. Note that the overdensity satisfies the inequality $-1 \le \delta < \infty$.

In this section, we will specialize in small perturbations, where $\delta \ll 1$.

Standard inflationary theory predicts that the distribution of the primordial density fluctuations are Gaussian. This means

$$P(\delta)d\delta = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\delta^2/(2\sigma^2)} d\delta. \tag{2}$$

From their initial size $\delta \simeq 10^{-5}$ these fluctuations later grow, as we described here, and remain Gaussian until the onset of nonlinearity at recent times when $\delta \gtrsim 1$.

Note too that, by definition, $\delta \equiv (\rho-\bar{\rho})/\bar{\rho}$ can be between $-1$ and $+\infty$. So, while the small perturbations $\delta \ll 1$ are symmetrical around zero (being Gaussian!), we *know* that large fluctuations $\delta \gtrsim 1$, cannot remain Gaussian, simply because large fluctuations in $\delta$ will be $\gg 1$, while the underdensities can never fall below $-1$.

However, the Gaussianity of the density fluctuations is not a fundamental physical principle, and testing it is well worthwhile. In the recent decade, tests of so-called non-Gaussianity (meaning, departures from the Gaussian distribution in Eq. (2) have become a hot topic in cosmology, both on the experimental front (measuring the distribution of overdensities) and theoretical one (finding that some non-standard, and perhaps more realistic, inflationary models predict non-Gaussianities of measurable magnitude). In what follows, however, we assume the standard picture of Gaussian perturbations at early times.

**Two point function for a point process.** Consider a point process (process with point particles in space) and let the mean density of points be $n$. Then the probability of finding a particle in an infinitesimal volume $dV$ is[1]

$$dP = ndV \tag{3}$$

Now consider the probability of finding two particles, one in volume $dV_1$ and another in volume $dV_2$; this is

$$dP = n^2(1 + \xi(r_{12}))\, dV_1 dV_2 \tag{4}$$

Here $\xi$ is the *excess* probability of finding the second particle a distance $r_{12}$ away (we assume isotropy here, so $\xi$ can depend at most on distance, and not direction). In other words, given that you observe particle 1 in $dV_1$, the probability that you find the second particle in $dV_2$ is

$$dP(2|1) = n(1 + \xi(r_{12}))\, dV_2 \tag{5}$$

For a pure Poisson ("random") process, there is no correlation between counts in volumes $dV_1$ and $dV_2$, so that $\xi = 0$ for a pure Poisson process. Measured correlation functions for galaxies and clusters have been some of the first aspects of LSS to be measured; to a good approximation

$$\boxed{\xi(r) = \left(\frac{r}{r_0}\right)^{\gamma}} \tag{6}$$

with $\gamma \approx -1.8$. The value of $r_0$ depends on the type of object we are talking about; for galaxies, $r_0 \simeq 5\, h^{-1}$Mpc, while for clusters $r_0 \simeq 20\, h^{-1}$Mpc. These rough scaling laws for the two point correlation functions and the associated $\gamma$ and $r_0$ values have already been known in the early 1980s.

Consider also the expected number of particles in the space of volume $V$, around a particle centered at that volume (and excluding that particle). The expected number is

$$\langle N \rangle = nV + n \int \xi(r) dV. \tag{7}$$

where the integral runs over the volume $V$ and $r$ is the distance from the central particle.

Finally, consider the *three-point correlation function* in real space $\zeta_{123}$. It can be defined via a probability of finding three particles in volumes $dV_1$, $dV_2$ and $dV_3$:

$$dP = n^3[1 + \xi(r_{12}) + \xi(r_{13}) + \xi(r_{23}) + \zeta(r_{123}))]\, dV_1 dV_2 dV_3 \tag{8}$$

**Continuous processes.** Consider now a continuous density field $\rho(\mathbf{x})$. Then the two point correlation function can be defined as

$$\boxed{\xi(r) = \frac{\langle\, [\, \rho(\mathbf{x}+\mathbf{r}) - \langle\rho\rangle\, ]\, [\, \rho(\mathbf{x}) - \langle\rho\rangle\, ]\, \rangle_{\mathbf{x}}}{\langle\rho\rangle^2} = \langle\, \delta(\mathbf{x}+\mathbf{r})\, \delta(\mathbf{x})\, \rangle_{\mathbf{x}}} \tag{9}$$

[1]Note that, strictly speaking, this is the number of points found in the volume. But since $dP \ll 1$, this is equivalent to the probability of finding a point particle.

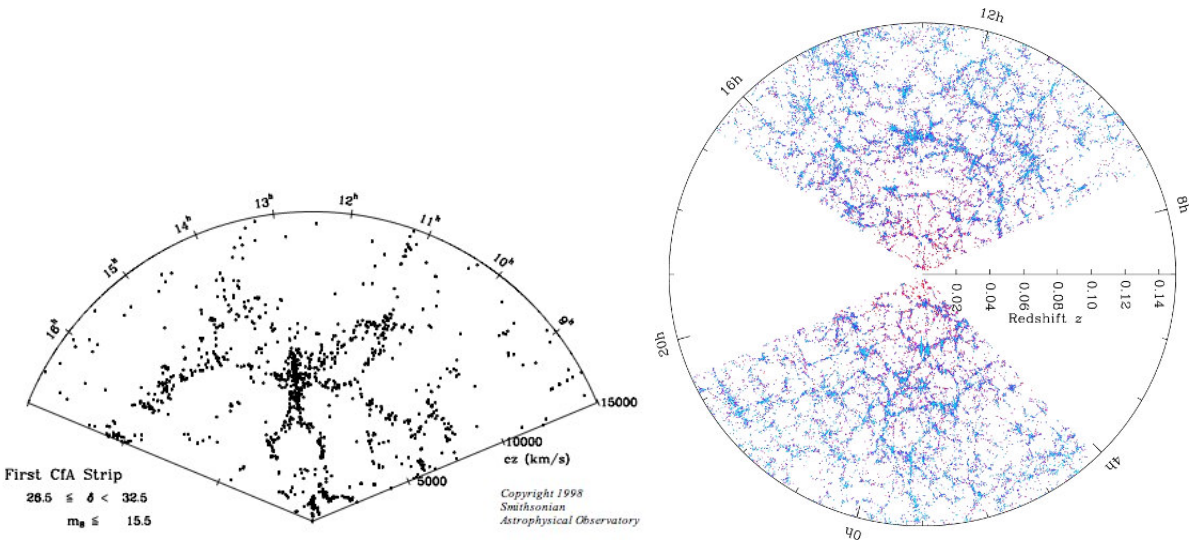


Figure 1: Left panel: Distribution of galaxies in the Cfa (Harvard Center for Astrophysics) galaxy redshift survey, with about 1100 galaxies. Notice the famous 'stick man' structure. Adopted from de Lapparent, Geller & Huchra, ApJ 302, L1 (1986). Right panel: Distribution of galaxies in the complete Baryon Oscillation Sky Survey (BOSS; extension of SDSS). Each point represents one of roughly a million galaxies with accurate redshift measurements. We are at the center of the diagram. Adopted from `http://www.sdss3.org/science`.

where the averaging is done over all $\mathbf{x}$, and which could be re-written as

$$\langle\, \rho(\mathbf{x}+\mathbf{r})\, \rho(\mathbf{x})\, \rangle_{\mathbf{x}} = \langle\rho\rangle^2\, [1+\xi(r)]. \tag{10}$$

We could similarly define the continuous version of the *three-point correlation function*

$$\boxed{\zeta(r,s,|\mathbf{r}-\mathbf{s}|) = \frac{\langle\, [\,\rho(\mathbf{x}+\mathbf{r})-\langle\rho\rangle\,]\,[\,\rho(\mathbf{x}+\mathbf{s})-\langle\rho\rangle\,]\,[\,\rho(\mathbf{x})-\langle\rho\rangle\,]\,\rangle_{\mathbf{x}}}{\langle\rho\rangle^3} = \langle\, \delta(\mathbf{x}+\mathbf{r})\, \delta(\mathbf{x}+\mathbf{s})\, \delta(\mathbf{x})\, \rangle_{\mathbf{x}}} \tag{11}$$

which could be re-written as

$$\langle\, \rho(\mathbf{x}+\mathbf{r})\, \rho(\mathbf{x}+\mathbf{s})\, \rho(\mathbf{x})\, \rangle_{\mathbf{x}} = \langle\rho\rangle^3\, [1+\xi(r)+\xi(s)+\xi(|\mathbf{r}-\mathbf{s}|)+\zeta(r,s,|\mathbf{r}-\mathbf{s}|). \tag{12}$$

**Density perturbations: Fluid equations.** Let us start with three fundamental equations that describe the evolution of a fluid. In what follows,

$$\frac{D}{Dt} \equiv \frac{\partial}{\partial t} + \mathbf{u}\cdot\nabla_{\mathbf{x}} \tag{13}$$

is the so-called *convective derivative* (or Lagrangian derivative, or total derivative), and measures what someone moving with the flow - whose velocity is $\mathbf{u}$ – would measure. In other words, the convective derivative is the derivative measured along the flow lines.

First, the continuity equation describes the mass conservation

$$\frac{D\rho}{Dt} + \rho(\nabla_{\mathbf{x}}\cdot\mathbf{u}) = 0 \quad \text{(continuity equation)}. \tag{14}$$

Then, the Euler equation correspond to the conservation of momentum, and are really three equations of motion (one for each direction)

$$\frac{D\mathbf{u}}{Dt} = -\frac{\nabla_{\mathbf{x}} p}{\rho} - \nabla_{\mathbf{x}}\Phi \quad \text{(Euler equation)}. \tag{15}$$

Finally, the Poisson equation relates the gravitational potential and (matter) density:

$$\nabla^2\Phi = 4\pi G\rho \quad \text{(Poisson, equation)}. \tag{16}$$

We would like to consider some of these equations in comoving coordinates. To that effect, note that the relation between a physical coordinate $\mathbf{x}$ and comoving coordinate $\mathbf{r}$ is, by definition[2]

$$\mathbf{x} = a(t)\mathbf{r} \tag{17}$$

so that the physical and comoving velocities are related as

$$\mathbf{u} = \dot{a}\mathbf{r} + a\dot{\mathbf{r}} = H\mathbf{x} + \mathbf{v} \tag{18}$$

where $\mathbf{v} \equiv a\dot{\mathbf{r}}$ is the peculiar velocity. This is of course just the familiar Hubble law, "corrected" for the peculiar velocity of the fluid element. It is therefore easy to see the identities for going from physical to comoving time derivative and gradient:

$$\nabla_{\mathbf{x}} \to \frac{1}{a}\nabla_{\mathbf{r}}; \quad \frac{\partial}{\partial t} \to \frac{\partial}{\partial t} - \frac{\dot{a}}{a}\mathbf{r}\cdot\nabla_{\mathbf{r}}. \tag{19}$$

With the help of these, one can rewrite the three equations as

$$\frac{\partial\delta}{\partial t} + \frac{1}{a}\nabla\cdot[(1+\delta)\mathbf{v}] = 0 \tag{20}$$

$$\frac{\partial\mathbf{v}}{\partial t} + \frac{\dot{a}}{a}\mathbf{v} + \frac{1}{a}(\mathbf{v}\cdot\nabla)\mathbf{v} = -\frac{\nabla\Phi}{a} - \frac{\nabla p}{a\bar{\rho}(1+\delta)} \tag{21}$$

$$\nabla^2\Phi = 4\pi G\bar{\rho}a^2\delta \tag{22}$$

where in the last equation we also identified $\Phi \to \Phi + a\dot{a}r^2/2$. Recall again that the partial derivative here is for fixed comoving location $\mathbf{r}$, and that the gradient is also with respect to $\mathbf{r}$.

**Thermodynamics.** The goal now is to rewrite the gradient of pressure term on the rhs of Eq. (21). To that effect, remember the first law of thermodynamics

$$TdS = dU + pdV \tag{23}$$

where $dS$ is the change in the entropy of the system, $dU$ is flow of energy into the system, and $pdV$ is work done on the system. Also

$$p = nkT = \frac{\rho kT}{\mu m_P} \tag{24}$$

for an ideal gas, where $\mu$ is mean molecular mass and $m_p$ is the proton mass. Finally for monoatomic gas, $U = (3/2)kT = (3/2)(p/\rho)$, so that the first law for a unit mass becomes

$$TdS = d\left(\frac{3}{2}\frac{p}{\rho}\right) + pd\left(\frac{1}{\rho}\right) \tag{25}$$

[2]Note that our convention is exactly the opposite from those in the Mo, van den Bosch, and White book. But ours is much better, since we definitely want $\mathbf{r}$ to be the comoving coordinate, as per widespread use.

Substituting temperature from Eq. (24) we then have

$$d\ln p = \frac{5}{3} d\ln\rho + \frac{2}{3}\frac{\mu m_p}{k}\, S d\ln S \tag{26}$$

which can be integrated to give

$$p \propto \rho^{5/3} \exp\left(\frac{2}{3}\frac{\mu m_p}{k}\, S\right) \tag{27}$$

Then we can express the gradient of pressure as

$$\frac{\nabla p}{\bar{\rho}} = \frac{1}{\bar{\rho}}\left[\left(\frac{\partial p}{\partial \bar{\rho}}\right)_S \nabla\rho + \left(\frac{\partial p}{\partial S}\right)_\rho \nabla S\right] \tag{28}$$

$$= c_s^2\, \nabla\delta + \frac{2}{3}(1+\delta) T\, \nabla S \tag{29}$$

where

$$c_s \equiv \left(\frac{\partial p}{\partial \rho}\right)_S^{1/2} \tag{30}$$

is the (adiabatic) speed of sound. With the application of Eq. (28), the Euler Eq. (21) becomes

$$\frac{\partial \mathbf{v}}{\partial t} + \frac{\dot{a}}{a}\mathbf{v} + \frac{1}{a}(\mathbf{v}\cdot\nabla)\mathbf{v} = -\frac{\nabla\Phi}{a} - \frac{c_s^2}{a}\frac{\nabla\delta}{(1+\delta)} - \frac{2T}{3a}\nabla S \tag{31}$$

Finally, in linear theory we can ignore the terms quadratic in $v$ to get

$$\frac{\partial \mathbf{v}}{\partial t} + \frac{\dot{a}}{a}\mathbf{v} = -\frac{\nabla\Phi}{a} - \frac{c_s^2}{a}\nabla\delta - \frac{2T}{3a}\nabla S \tag{32}$$

**Curl modes decay.** In linear theory, and curl modes in the velocity field decay as the scale factor. This can be seen by operating with $\nabla\times$ on both sides of Eq. (32):

$$\left[\frac{\partial}{\partial t} + \frac{\dot{a}}{a}\right](\nabla\times\mathbf{v}) = 0 \tag{33}$$

from it follows that

$$(\nabla\times\mathbf{v}) \propto \frac{1}{a}. \tag{34}$$

Therefore, even if there are initial curl modes in the velocity distribution of baryonic or dark matter particles in the universe, they decay in time, becoming soon negligible. This "basically comes from the conservation of angular momentum in the expanding universe" (Mo, van den Bosch and White). This fact plays an important role in at least one strategy for reconstructing the large-scale velocity field in the universe.

**Temporal evolution of fluctuations: general case.** Taking the derivative of the linear Euler equation (32) and combining with the continuity equation (20) and Poisson Eq. (22), it is easy to get the general equation for the evolution of fluctuations:

$$\frac{\partial^2\delta}{\partial t^2} + 2\frac{\dot{a}}{a}\frac{\partial\delta}{\partial t} = 4\pi G\bar{\rho}\delta + \frac{c_s^2}{a^2}\nabla^2\delta + \frac{2}{3}\frac{T}{a^2}\nabla^2 S \tag{35}$$

This is a second-order ordinary differential equation, which means there is a growing and decaying solution – of course, the former will dominate over time. The key source term is the first term on the right-hand side, which ensures that perturbations grow due to gravitational instability. Finally, the second term on the lhs is the friction term, which causes fluctuations grow slower than they would in a static universe: instead of exponentially in time, the perturbations grow as a power-law.

To make even better progress, people usually expand the overdensity in Fourier space:

$$\delta_{\mathbf{k}}(t) = \frac{1}{\sqrt{V}} \int \delta(\mathbf{r}, t) e^{-i\mathbf{k}\cdot\mathbf{r}} d^3\mathbf{r} \tag{36}$$

where $V_u$ is the volume of the larger region over which the perturbations are assumed to be periodic. Note that $\mathbf{k}$ and $\mathbf{r}$ are both comoving quantities.

The rule of thumb of Fourier-transforming an equation is that the time derivatives remain unchanged, while the gradients change as $\nabla \to -i\mathbf{k}$ and $\nabla^2 \to -k^2$. Then the equation above:

$$\boxed{\frac{\partial^2 \delta_{\mathbf{k}}}{\partial t^2} + 2\frac{\dot{a}}{a}\frac{\partial \delta_{\mathbf{k}}}{\partial t} = \left(4\pi G\bar{\rho} - \frac{k^2 c_s^2}{a^2}\right)\delta_{\rm k} - \frac{2}{3}\frac{T}{a^2}k^2 S_{\mathbf{k}}} \tag{37}$$

**Isentropic (adiabatic) and isocurvature initial conditions.** There exist two general kinds of initial conditions (ICs) that a universe could, in principle, have. These are

1. Isentropic ICs, where there is no fluctuation in the entropy in the initial conditions, so that $\nabla S = 0$ (hence, isentropic). Confusingly, these fluctuations are most often called adiabatic, which strictly speaking is actually the $\dot{S} = 0$ condition, not $\nabla S = 0$. With the isentropic ICs fluctuations in various components (matter, radiation, neutrinos etc) are proportional to each other. Here, but there are fluctuations in curvature. This is the kind of initial condition that inflation predicts, and that current data favor. Note the entropy is given by the ratios of number densities of species, for example $s = n_B/n_\gamma$, so that the isentropic condition is that $\delta n/n$ be the same for each species. For baryons or CDM, $\delta n/n = \delta\rho/\rho$ since $\rho = n \cdot m$ ($m$ being particle mass). For photons or relativistic neutrinos, recall that $n \propto T^3$ and $\rho \propto T^4$, so that $\delta n/n = (3/4)(\delta\rho/\rho)$. Then the adiabatic initial condition becomes, in terms of number density fluctuations

$$\frac{1}{3}\delta_B = \frac{1}{3}\delta_c = \frac{1}{4}\delta_\gamma = \frac{1}{4}\delta_\nu \tag{38}$$

   where $\delta \equiv \delta\rho/\rho$ and the indices refer, respectively, to baryons, cold dark matter, photons, and relativistic neutrinos.

2. Isocurvature ICs, where there is identically zero curvature fluctuation, so that the fluctuations in different species effectively sum to zero. In this case, entropy modulation is generated. There is a number of possibilities, which correspond to which of the equalities in Eq. (38 is violated, for example one mode corresponds to

$$S = \frac{\delta\rho_m}{\rho_m} - \frac{3}{4}\frac{\delta\rho_\gamma}{\rho_\gamma} \tag{39}$$

   where $\rho_m = \rho_B + \rho_c$.

Current cosmological constraints indicate that fluctuations are isentropic, and only a very small admixture of isocurvature fluctuations (ballpark 1%, maximum) is allowed by the data. In what follows, we assume isentropic fluctuations only, and ignore the entropy term in Eq. (37).

**Gravitational Instability.** Assume from now on

- isentropic (lazily often called adiabatic) fluctuations with zero entropy fluctuation;
- drop the subscript $\mathbf{k}$, remembering that $\delta$ is referring to the Fourier component now.

Equation (35) can be written as

$$\frac{\partial^2 \delta_{\mathbf{k}}}{\partial t^2} + 2\frac{\dot{a}}{a}\frac{\partial \delta_{\mathbf{k}}}{\partial t} = -\omega^2 \delta \tag{40}$$

where

$$\omega^2 = \left(\frac{k^2 c_s^2}{a^2} - 4\pi G\bar{\rho}\right) \equiv (k^2 - k_J^2)\frac{c_s^2}{a^2} \tag{41}$$

which defines a characteristic *Jeans length*

$$\lambda_J \equiv \frac{2\pi a}{k_J} \equiv c_s\sqrt{\frac{\pi}{G\bar{\rho}}} \tag{42}$$

The solution *in the case of a static universe* (which ignores the $\dot{\delta}$ term) is

- Oscillating, $\delta(t) \propto \exp(\pm i\omega t)$ for $\lambda < \lambda_J$ (or $k > k_J$, that is when $\omega^2 > 0$)
- Exponentially growing, $\delta(t) \propto \exp(|\omega| t)$ for $\lambda > \lambda_J$ (or $k < k_J$, that is when $\omega^2 < 0$)

For the expanding universe a qualitatively similar results holds: only fluctuation Fourier modes larger than the Jeans length $\lambda_J$ grow – though as a power-law, not exponentially. In other words, only objects more massive than the Jeans mass can form, where this mass is defined as

$$M_J \equiv (4\pi/3)(\lambda_J/2))^3 \bar{\rho} \sim c_s^3\, \bar{\rho}^{-1/2}. \tag{43}$$

It turns out that Jeans mass is large before recombination, but drops like a rock at this time. The reason is that the speed of sound goes from somewhat relativistic, due to the coupling of baryons to photons before recombination, to near zero when the baryons are decoupled. In more detail, before recombination photons are baryons are tightly coupled, and the speed of sound is close to that of radiation alone, which is $c/\sqrt{3}$

$$(c_s)^{\text{before recomb}} = \frac{c}{\sqrt{3\left(1 + \dfrac{3\rho_B(z)}{4\rho_\gamma(z)}\right)}} \simeq \frac{c}{\sqrt{3}} \tag{44}$$

where the coefficient 3/4 comes about because under adiabatic compression, the energy densities change with volume as $\rho_M \propto V^{-1}$ and $\rho_\gamma \propto V^{-4/3}$. Then it follows

$$(M_J)^{\text{before recomb}} \simeq O(10^{19} M_\odot) \tag{45}$$

which corresponds to mass larger than that of any known object in the universe. Before recombination, therefore, pressure forbids gravitational collapse of structures.

*After* recombination, however, baryons are not coupled to photons any more and they can be considered a non-relativistic monoatomic gas with speed of sound

$$(c_s)^{\text{after recomb}} = \sqrt{\frac{5kT}{3m_p}} \tag{46}$$

which evaluates to about $6.5\,\mathrm{km\,s^{-1}}$ so that

$$(M_J)^{\text{after recomb}} \simeq O(10^5 M_\odot) \tag{47}$$

and further falls from that value because $c_s \propto T^{1/2}$. Therefore, after recombination pressure support does not block the formation of cosmologically interesting structures.

**Linear growth equation.** On scales smaller than the horizon (so that relativistic effects do not apply) and for matter at late times (so that the speed of sound is negligible), Eq. (35) reads

$$\boxed{\ddot{\delta} + 2H\dot{\delta} - 4\pi G\rho_M(t)\delta = 0}. \tag{48}$$

Note that growth equation, being 2nd order ODE, has two solutions, growing and decaying. We are obviously interested in the growing solutions (the decaying one becomes unimportant quickly).

**Solutions to the growth equation.** Let us work out solution to the growth equation in a few simple cases.

- *Radiation-dominated era.* In the RD era, recall that $a \propto t^{1/2}$ so that $H(t) \equiv \dot{a}/a \propto 1/(2t)$. Also, note that the last term is negligible since the matter density is negligible (so $4\pi G\rho_M \ll H^2$). Therefore, we need to solve the equation

  $$\ddot{\delta} + 2H\dot{\delta} = 0 \tag{49}$$

  whose solution is

  $$\boxed{\delta(t) = A_1 + A_2 \ln t \qquad \text{(radiation dominated)}.} \tag{50}$$

  Therefore, perturbations in RD universe growth extremely slowly (logarithmically) — basically they don't grow. This will become important in a bit.

- *Matter-dominated era.* In the MD era, recall that $a \propto t^{2/3}$ so that $H(t) \equiv \dot{a}/a \propto 2/(3t)$. Also, in the MD era the Hubble parameter is dominated by matter density, so that $4\pi G\rho_M = (3/2)H^2$. Let us assume that $\delta(t) \propto t^n$; then the growth equation simplifies to

  $$n(n-1) + \frac{4}{3}n - \frac{2}{3} = 0 \tag{51}$$

  whose solutions are easy to obtain: $n = +2/3$ and $-1$, so that

  $$\boxed{\delta(t) = B_1 t^{2/3} + B_2 t^{-1} \qquad \text{(matter dominated)}.} \tag{52}$$

  Since you can recall that $a(t) \propto t^{2/3}$ in the MD era, in the matter-dominated era the (growing mode of) the perturbations grow proportional to the scale factor

  $$\boxed{\delta(t) \propto a(t) \qquad \text{(matter dominated)}}. \tag{53}$$

- *Lambda-dominated era.* In the LD era, which will presumably happen in the future when dark energy dominates the universe completely, the scale factor grows exponentially $a \propto e^{Ht}$[3] so that $H(t) \equiv H_\Lambda = \text{const.}$. Also, note that the last term is negligible since the matter density is negligible relative to vacuum energy. Therefore, we need to solve the equation

$$\ddot{\delta} + 2H_\Lambda \dot{\delta} = 0 \tag{54}$$

whose solution is

$$\boxed{\delta(t) = C_1 + C_2 e^{-2H_\Lambda t} \qquad \text{(Lambda dominated).}} \tag{55}$$

Therefore, density perturbations do not grow in Lambda-dominated universe; they reach a finite size and stay there.

**Growth of perturbations: more general solutions.** A more general solution to the growth equation can be found, and it's valid for the case when dark energy is the cosmological constant (i.e. when $w(z) = -1$), but *not* for more general arbitrary values of the equation of state. First, we helpfully define the growth function as

$$\boxed{D(a) = \delta(a)/\delta(1)} \qquad \text{(growth function definition).} \tag{56}$$

where clearly $D$ at the present time is unity, $D(1) = 1$.

For this case, one can show that the growing solution is

$$D(a) \propto \frac{H(a)}{H_0} \int_0^a \frac{da'}{\left[\Omega_M/a' + \Omega_\Lambda a'^2 + (1 - \Omega_M - \Omega_\Lambda)\right]^{3/2}} \quad \text{(only for } w = -1) \tag{57}$$

where the constant of proportionality is determined by requiring that $D(a = 1) = 1$.

Finally, the most general solution, valid for any equation of state $w(z)$ of dark energy, requires solving the 2nd order ODE, except it can be written in a nice dimensionless form as

$$\boxed{2\frac{d^2 g}{d\ln a^2} + \left[5 - 3w(a)\Omega_{\rm DE}(a)\right]\frac{dg}{d\ln a} + 3\left[1 - w(a)\right]\Omega_{\rm DE}(a) g = 0} \tag{58}$$

where $g(a)$ is the "growth suppression factor" – that is, growth relative to that in Einstein-deSitter Universe. It is related to $D(a)$ via

$$D(a) \equiv \frac{a g(a)}{g(1)}. \tag{59}$$

In the best-fit ΛCDM cosmology ("benchmark model" from Ryden), the value of the suppression factor today is $g(1) \simeq 0.76$, while obviously $D(1) \equiv 1$ by definition.

**Power spectrum.** What is the distribution of $\delta$ in the universe? Does it have a lot of structure on small scales like the surface of a sandpaper, or on large scales like rolling hills? This question is best answered in Fourier space, looking at Fourier components of overdensity, $\delta_k$.

The overdensity can be expressed in its Fourier components in some comoving volume $V$ as

$$\delta(\vec{r}) = \frac{\sqrt{V}}{(2\pi)^3} \int \delta_{\vec{k}}\, e^{-i\vec{k}\vec{r}}\, d^3k \tag{60}$$

[3] One way to realize this without deriving anything is to remember that this is the scaling during inflation, when the universe is essentially vacuum-energy dominated.

and the Fourier components are, conversely

$$\delta_{\vec{k}} = \frac{1}{\sqrt{V}} \int \delta(\vec{r})\, e^{i\vec{k}\vec{r}}\, d^3r. \tag{61}$$

We still work in the limit $\delta(\vec{r}) \ll 1$ and $\delta_{\vec{k}} \ll 1$. Then, for example, each Fourier component satisfies the growth equation (48). Note too that $\delta_{\vec{k}}$ are complex numbers in general, and that their units are (distance)$^{-3/2}$ (while, of course, the real-space version is dimensionless).

If we shift $\vec{r}$ by some $\Delta\vec{r}$, then

$$\delta_{\vec{k}} \to \delta_{\vec{k}}\, e^{i\vec{k}\Delta\vec{r}} \tag{62}$$

and the two point function transforms as

$$\langle \delta_{\vec{k}} \delta^*_{\vec{k}'} \rangle \to e^{i(\vec{k}-\vec{k}')\Delta\vec{r}} \langle \delta_{\vec{k}} \delta^*_{\vec{k}'} \rangle \tag{63}$$

Due to homogeneity, we know that this quantity must not depend on $\Delta\vec{r}$. Therefore, the quantity must be proportional to $\delta(\vec{k}-\vec{k}')$, and the remaining dependence is only on $\vec{k}$. Finally, also due to homogeneity, only the magnitude of the wavenumber, $k \equiv |\vec{k}|$, matters. Thus

$$\boxed{\langle \delta_{\vec{k}}\, \delta^*_{\vec{k}'} \rangle = (2\pi)^3\, \delta^{(3)}(\vec{k}-\vec{k}')\, P(k)} \tag{64}$$

where $P(k)$ is the *power spectrum* — the Fourier transform of the 2-point function. Note that $P(\vec{k}) = P(k)$ due to homogeneity.

The power spectrum is essentially defined as the ensemble average, over all universes, of the square of the Fourier component $\delta_{\vec{k}}$. Since we are usually not able to average over different universes, we average in our universe over locations.

The power spectrum tells us how much power is on different scales — that is, different wavenumbers $k$. For example, imagine a completely unrealistic case when the sky looks like a chessboard, with white and black pixels. Let the size of the pixel be $R$. Then the power spectrum will be zero at all scales, except at $k_r = 2\pi/R$, where it will peak. At smaller scales you are looking within a pixel where there is no variation in color, while at much larger scales, you are averaging over many pixels and get washed out signal. It's only at scale $R$ (or $k_R$, in Fourier space), that you see black-white and thus have nonzero power.

**Inflationary prediction for the shape of $P(k)$.** In our universe, however, inflation generates power on all scales. But what is the relative power on different $k$? A conjecture was given by Harrison, Zel'dovich, and Peebles (all working independently) in the late 1960s, that

$$\boxed{P(k) \propto k^n \quad \text{with} \quad n \simeq 1} \qquad \text{(Harrison, Zel′dovich, Peebles spectrum)} \tag{65}$$

If $n$ were much bigger than 1, there would be too much small-scale power (large $k$), and too many black holes would be created too early. If $n$ were much less than one, there would be too little power on small scales relative to large scales, and huge superclusters and voids would dominate over the much smaller galaxies, which is not what we observe.

Two remarkable things have happened since the Harrison-Zel'dovich-Peebles conjecture circa 1969

- Inflation, proposed in 1980 by Guth, predicts that $n = 1 - 6\epsilon + 2\eta + O(\epsilon^2, \eta^2)$, where $\epsilon$ and $\eta$ are the so-called slow-roll parameters and are related to the first two derivatives of the effective potential of inflaton wrt the field value. Thus, inflation is right on the HZP prediction for $n$.

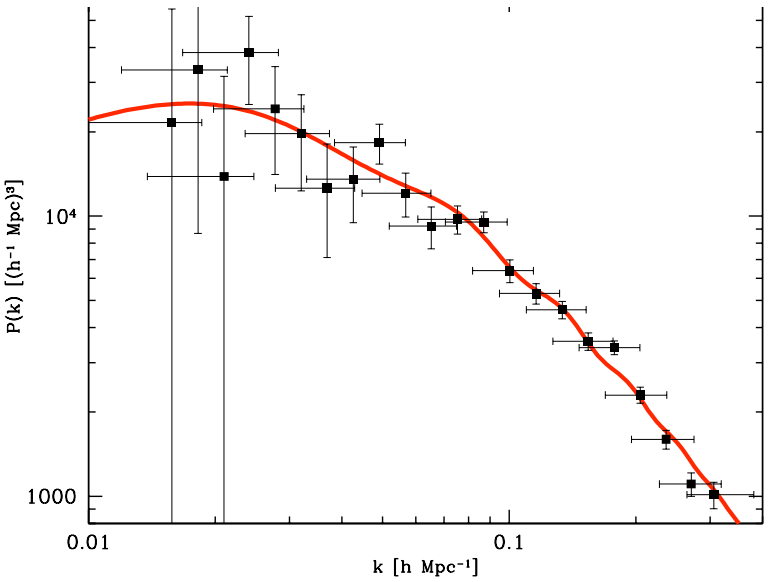


Figure 2: Power spectrum of galaxies from the Sloan Digital Sky Survey (Tegmark et al, 2004). The measurements of the spectrum have been further brought into a new basis so that they are 100% uncorrelated from each other — don't worry about that detail here. Note the small wiggles in the theoretical prediction — the baryon acoustic oscillations (they too have been detected in the data a couple of years after this paper).

- $n$ has actually been measured by modern experiments like Planck to be just below one; $n = 0.963 \pm 0.005$. This fits the inflationary prediction right on target, since usually (but by no means always!) $-6\epsilon + 2\eta$ is slightly negative as predicted by inflationary models.

**Relation between $P(k)$ and $\xi(r)$** In particular, we can relate the correlation function to the power spectrum. First, recall that the correlation function is given by

$$\boxed{\langle \delta_{\vec{r}_1} \delta^*_{\vec{r}_2} \rangle = \xi(r_{12})} \quad \text{where } r_{12} = |\vec{r}_1 - \vec{r}_2| \tag{66}$$

where again, due to homogeneity, the correlation function must depend only on the distance between the two vectors.

Now we can compare the two via

$$\begin{aligned} P(k) &= \langle \delta_{\vec{k}}\, \delta^*_{\vec{k}} \rangle \ (= \langle \delta_{\vec{k}}\, \delta_{-\vec{k}} \rangle) && (67)\\ &= \frac{1}{V} \int \int \langle \delta(\vec{r}_1) \delta^*(\vec{r}_2) \rangle e^{-i\vec{k}\vec{r}_1} e^{i\vec{k}\vec{r}_2} d^3\vec{r}_1\, d^3\vec{r}_2 && (68)\\ &= \int \xi(r_{12}) e^{-i\vec{k}\vec{r}_{12}} d^3\vec{r}_{12} && (69)\\ &= \int \xi(r) e^{-i\vec{k}\vec{r}} d^3\vec{r} && (70) \end{aligned}$$

where in going from the second to third line we switched to $(\vec{r}_{12}, \vec{r}_2)$, and trivially integrated

over $\vec{r}_2$ and used $\int d^3\vec{r}_2 = V$. We can further make progress on this integral by

$$P(k) = \int_0^\infty \xi(r) \left( \int_0^{2\pi} \int_0^\pi e^{-ikr\cos\theta} \sin\theta d\theta d\phi \right) r^2 dr \tag{71}$$

$$= 2\pi \int_0^\infty \xi(r) \left( \int_1^{-1} e^{-ikru}(-du) \right) r^2 dr \tag{72}$$

The integral over $\cos\theta$ evaluates to $2\sin(kr)/(kr)$, so the result is

$$P(k) = \frac{4\pi}{k} \int_0^\infty \sin(kr)\xi(r) r dr. \tag{73}$$

Notice a few nice things

- the power spectrum does not depend on the volume $V$, so we can meaningfully talk about the power spectrum independent of the choice of $V$
- The power spectrum is a Fourier transform of the 2-point correlation function
- The units of $P(k)$ are just those of (comoving) volume, so that $P(k)/V$ is dimensionless, or else $k^3P(k)$ is dimensionless.

Now let us inverse-Fourier transform to get

$$\xi(r) = \frac{1}{(2\pi)^3} \int P(k) e^{i\vec{k}\vec{r}} d^3\vec{k} = \frac{1}{2\pi^2 r} \int_0^\infty P(k) \sin(kr) k dk. \tag{74}$$

Let us now evaluate the *zero-lag* correlation function $\xi(r = 0)$

$$\xi(0) = \frac{1}{2\pi^2} \int_0^\infty P(k) \lim_{r\to 0} \frac{\sin(kr)}{kr} k^2 dk \tag{75}$$

$$\equiv \int_0^\infty \Delta^2(k) d\ln k \tag{76}$$

where we have defined the *logarithmic band power*

$$\boxed{\Delta^2(k) \equiv \frac{k^3 P(k)}{2\pi^2}} \tag{77}$$

which is the contribution to variance per log wavenumber. If the peak of $\Delta^2(k)$ is at some $k_*$, then fluctuations in $\delta$ are dominated by wavelengths of order $2\pi/k_*$. Notice that this logarithmic band power (sometimes also referred to as the power spectrum) is dimensionless.

Note also that the integral in Eq. (76) is badly divergent in the ultraviolet (large $k$ limit). What is the intuitive reason for this? At any rate, this already indicates that we will have to *smooth* the density field before calculating the variance (or covariance). More a bit further below.

**Specialness of the Peebles-Harrison-Zeldovich spectrum.** Consider an object of (comoving) size $r$ and mass $M \propto r^3$. One can approximately integrate Eq. (75), with limits from $k = 0$ to $k = 1/r$, at which point $\sin(kr)/kr$ begins to wildly oscillate. Doing this gives $\xi(r) \propto \int_0^{1/r} k^{n+2} dk \propto r^{-(n+3)}$. Then the mass rms fluctuation is

$$\delta_{M,\mathrm{rms}} \equiv \left\langle \left( \frac{\delta M}{M} \right)^2 \right\rangle^{1/2} \propto r^{-(n+3)/2} \propto M^{-(n+3)/6}, \tag{78}$$

where the last equality follows from $M \propto r^3$. Note already one special result: for $n = 0$, we obtain the white-noise power spectrum. This follows because $\delta_{M,\mathrm{rms}} \propto M^{-1/2}$, similarly as in the familiar case of $\delta N/N \propto N^{-1/2}$, signifies Poisson noise type spectrum of fluctuations.

Inflation and observations instead favor something like $n \simeq 1$. For $n = 1$ exactly, we have the so-called scale-invariant spectrum where every horizon-scale fluctuation mode has the same amplitude. We can see this as follows: consider first the growth of *super*-horizon fluctuations in the radiation-dominated regime. From the Poisson equation, $-k^2\Phi/a^2 \propto \rho\delta$, it follows that in the radiation dominated epoch, when $\rho \propto a^{-4}$, keeping $\Phi$ constant at super-horizon scales (remember the constancy of the Bardeen potential) requires $\delta \propto a^2$. Similarly, in the matter-dominated epoch, $\rho \propto a^{-3}$ and then keeping $\Phi$ constant at super-horizon scales requires $\delta \propto a$. Now the growth of $\delta_{M,\mathrm{rms}}$ during the two regimes, is the *same* in terms of the Hubble distance as it turns out:

$$\begin{cases} \mathrm{RD}\,(\delta \propto a^2): & k_H = aH \propto a \times a^{-2} \propto a^{-1} \Longrightarrow R_H \propto a \Longrightarrow \delta \propto R_H^2 \\ \mathrm{MD}\,(\delta \propto a): & k_H = aH \propto a \times a^{-3/2} \propto a^{-1/2} \Longrightarrow R_H \propto a^{1/2} \Longrightarrow \delta \propto R_H^2 \end{cases} \tag{79}$$

In either case $\delta \propto R_H^2$. The mass enclosed in the Hubble sphere is $M \propto R_H^3$, and thus

$$\delta_{M,\mathrm{rms}} \propto R_H^2 M^{-(n+3)/6} \propto M^{2/3} M^{-(n+3)/6} \to \mathrm{const} \qquad (\text{for } n = 1). \tag{80}$$

This justifies the usual statement that the amplitude of fluctuations of modes when they enter the horizon is constant for the PHZ spectrum of $n = 1$.

**Machine-friendly power spectrum.** Coding the power spectrum, particularly the dimensionless form in Eq. (77), is very useful. Here we show the formula for the power spectrum of dark matter density perturbations in standard FRW cosmology

$$\boxed{\Delta^2(k,a) = A\,\frac{4}{25}\frac{1}{\Omega_M^2}\left(\frac{k}{k_{\mathrm{piv}}}\right)^{n-1}\left(\frac{k}{H_0}\right)^4 [ag(a)]^2\,T^2(k)\,T_{\mathrm{nl}}(k)} \tag{81}$$

where

- $A$ is the normalization of the power spectrum (for the concordance cosmology, $A = 2.43 \times 10^{-9}$)
- $k_{\mathrm{piv}}$ is the "pivot" around which we compute the spectral index; for WMAP $k_{\mathrm{piv}} = 0.002\,\mathrm{Mpc}^{-1}$ is used (*beware — occasionally* $k = 0.05\,h\,\mathrm{Mpc}^{-1}$ *is used too, which is actually closer to the true pivot and anyway changes which amplitude* $A$ *is appropriate*)
- $[ag(a)]$ is the linear growth of perturbations. Note that in the EdS model $g(a) = 1$ identically and at all times, and in ΛCDM model $g(a)$ at recent times drops, down to the value of $\approx 0.76$ at $a = 1$. Note that $ag(a)$ is related to the growth function $D(a)$ via

$$\boxed{D(a) \equiv \frac{ag(a)}{g(1)},} \tag{82}$$

  so that $D(1) = 1$.
- $T(k)$ is the transfer function (see below).
- $T_{\mathrm{nl}}$ is prescription for a *nonlinear* power spectrum, which is usually calibrated from N-body simulations.

**So what is transfer function?** The transfer function encodes the growth of density fluctuations in the regimes when the universe is radiation or matter dominated.

Inflation predicts that fluctuations 'enter the horizon' (i.e. $\lambda < H^{-1}$) with the same amplitude. Larger wavelength (smaller $k$) fluctuations enter the horizon at later times. For example, fluctuations that are entering the horizon today are of wavelength of order $\lambda \sim H_0^{-1}$.

Consider the universe that were always matter dominated (MD). Then, even though longer wavelength fluctuations enter the horizon later, their amplitude would be the same[4] according to inflation. In that case, we would simply have $T(k) = 1$.

However, things are more complicated with the existence of the radiation-dominated (RD) era. Recall from above, during RD, perturbations don't grow (or only grow logarithmically). So, growth perturbation modes whose wavelength was small enough to have entered during the RD era was 'stunted', and they couldn't grow until the universe became MD! Conversely, the model whose wavelength was longer, and which entered the horizon during the MD era, never experienced the stunting of the growth.

We now derive the most prominent feature of the transfer function. Recall from the discussion in and around Eq. (79) that the super-horizon fluctuation $\delta$ scales with $R_H^2$ in both the radiation- and matter-dominated regime. Perturbation that enters the horizon at some scales $k > k_{\rm eq}$, where $k_{\rm eq}$ is the wavenumber corresponding to horizon scale at matter-radiation equality, will see its growth "stunted" relative to some other scale (e.g. $k_{\rm eq}$ itself) by

$$\frac{\delta_k}{\delta_{k_{\rm eq}}} = \left(\frac{R_{\rm H,k}}{R_{\rm H,eq}}\right)^2 \propto \left(\frac{k}{k_{\rm eq}}\right)^{-2}. \tag{83}$$

The growth suppression therefore goes as the square of the wavenumber; the higher the wavenumber is, the earlier the mode entered the horizon before equality, and the more its growth was stunted. Summarizing, the transfer function is

$$T(k) = \begin{cases} 1 & k \ll 1/L_0 \\ (kL_0)^{-2} & k \gg 1/L_0 \end{cases} \tag{84}$$

where $L_0$ is the characteristic scale: *size of the horizon and matter-radiation equality.* One can easily show that

$$L_0 \approx 12\,(\Omega_M h^2)^{-1}\,{\rm Mpc} \simeq 100\,{\rm Mpc}. \tag{85}$$

Therefore, the power spectrum $P(k)$ has the following asymptotic behaviors

$$P(k) \propto \begin{cases} k^n & k \ll 1/L_0 \\ k^{n-4} & k \gg 1/L_0 \end{cases} \tag{86}$$

Transfer functions can be inferred from the fits to numerical solutions, or else the exact output out from Einstein-Boltzmann solvers such as CAMB or CMBFAST.

**Two stories of structure formation.** It has a long time ago been recognized that there are two possible structure formation histories, depending on the *nature of dark matter*:

- **Cold dark matter (CDM).** Here dark matter is "cold", that is, *non-relativistic* at the time of matter-radiation equality (when, recall, the perturbations first get a chance to grow appreciably). An example of a CDM candidate is a WIMP (say, a supersymmetric particle such as the neutralino), or any other massive particle.

[4]The amplitude would be *nearly* the same; departures of this scale invariance are proportional to $n-1$, where $n$ is the spectral index (and recall, measurements show that $n \simeq 0.96$).

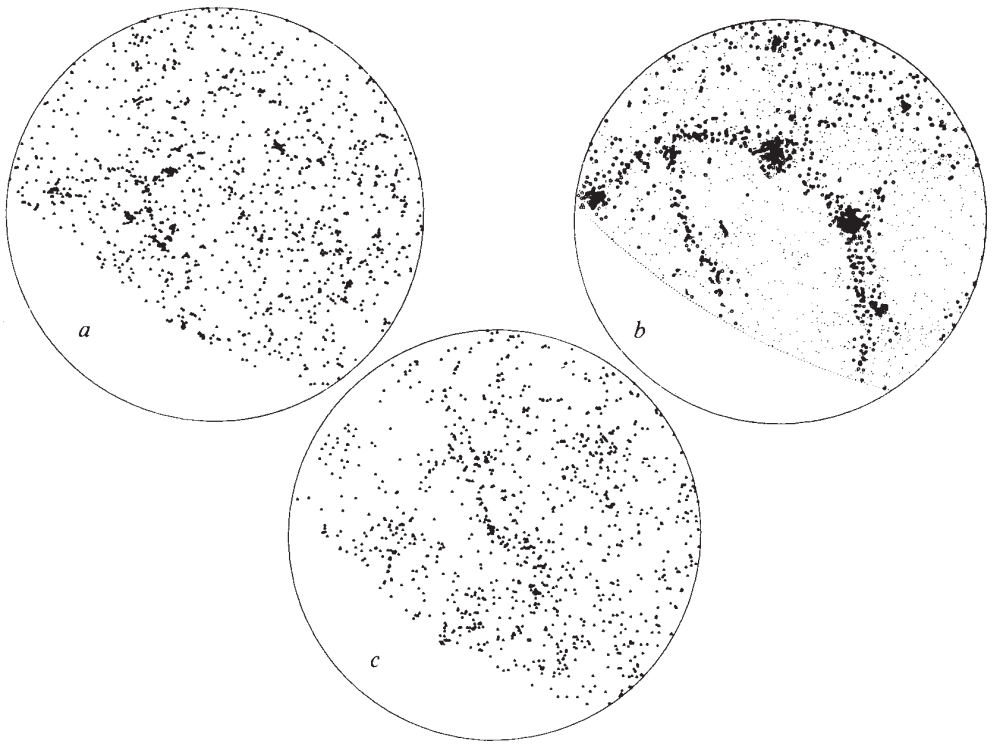


FIG. 2 Equal-area projections of the galaxy distribution on the northern sky and in artificial catalogues made from *N*-body simulations. *a*, Artificial catalogue in a standard CDM model with biased galaxy formation. *b*, Artificial catalogue in a neutrino-dominated model in which galaxy formation began at a redshift of 2.5; dots represent the neutrino distribution, circles represent the galaxies. *c*, Galaxies in the Center for Astrophysics northern survey. The outer circle represents galactic latitude +40°, and the empty regions lie at declinations below 0°.





Figure 3: Simulations that, in the late 1980s, showed that structure in the Cold Dark Matter dominated universe (top left) looks a lot more like measurements of galaxy distribution (bottom) than the neutrino-dominated universe (top right). In other words, the observed structure is consistent with the bottom-up structure formation scenario where the largest structures form last (such as that of the CDM) rather than the top-down formation where the largest structures form first. Adopted from White, Navarro, Evrard and Frenk, Nature, 366, 429 (1993).

- **Hot dark matter (HDM).** Here dark matter is hot, that is, *relativistic* at the time of matter-radiation equality. An example of an HDM candidate is a neutrino with mass of order a few electron-volts.

It turns out that hot dark matter does not clump very well – basically because the particles are relativistic – and all structures below a so-called free-streaming scale are washed out. This scale corresponds to something like $10^{15} M_\odot$. Therefore, the HDM scenario is 'top-down', since the largest objects (clusters) form first, and smaller objects form later.

In contrast, in the CDM scenario, the free-streaming scale is very small, and objects can grow just fine. This scenario is 'bottom-up' in that the least massive objects (stars, then galaxies) form first.

Cosmological observations clearly favor the CDM paradigm.

**Filtered density fields.** Often we would like to smooth the density field over some "window" in distance. In practice, in fact, the actual density field is grainy (think stars in a galaxy, galaxies in a cluster, etc) and the theory can only predict statistics for a smooth density field.

Formally, we define a smoothing function $W(r, R)$ where $r$ is the dependent variable, and $R$

is the characteristic smoothing scale. Popular choices are

$$W_{\rm G}(r,R) = \frac{1}{(2\pi)^{3/2}R^3}e^{-r^2/(2R^2)} \quad \text{(Gaussian smoothing)} \tag{87}$$

$$W_{\rm TH}(r,R) = \frac{1}{(4\pi/3)R^3}H(R-r) \quad \text{(Top}-\text{Hat smoothing)} \tag{88}$$

where $H(x)$ is the Heaviside step function; $H(x)=1$ for $x>0$ and $H(x)=0$ for $x<0$.

The smoothing is actually a convolution in real space, so that the smoothed density field becomes

$$\delta(\vec{r},R) = \int W(|\vec{r}-\vec{r}'|)\delta(\vec{r}')d^3\vec{r}' \tag{89}$$

Thankfully, convolution in real space corresponds to *multiplication* in Fourier space, so that

$$\delta_{\vec{k}}(R) = W(k,R)\delta_{\vec{k}} \tag{90}$$

and thus the power spectrum is

$$P(k,R) = |W(k,R)|^2 P(k) \tag{91}$$

The Fourier transforms are easily computed:

$$W_{\rm G}(k,R) = e^{-k^2R^2/2} \quad \text{(Gaussian)} \tag{92}$$

$$W_{\rm TH}(k,R) = 3\frac{\sin(kR)-kR\cos(kR)}{(kR)^3} = \frac{3j_1(kR)}{kR} \quad \text{(Top}-\text{Hat)} \tag{93}$$

where $j_1(x)$ is the spherical Bessel function of order one.

**Amplitude of mass fluctuations.** Let us adopt the most commonly used top-hat window/filter. What is the autocorrelation function $\xi(0,R)$? Well, going back to Eq. (76)

$$\xi_{\rm TH}(0) = \int_0^\infty \Delta^2(k)|W_{\rm TH}(k,R)|^2 d\ln k \tag{94}$$

or, renaming this quantity to agree with the literature

$$\boxed{\sigma^2(R) = \int_0^\infty \Delta^2(k)\left(\frac{3j_1(kR)}{kR}\right)^2 d\ln k} \tag{95}$$

This is the *rms amplitude (squared) of mass fluctuations* smoothed over scale R – a very important quantity in cosmology.

Historically, cosmologists have first studied clustering on galaxies on scales on about the size of a galaxy cluster (5-10 $h^{-1}$Mpc). In fact, an important quantity to choose was for $R = 8\,h^{-1}$Mpc:

$$\boxed{\sigma_8 \equiv \sigma(R = 8\,h^{-1}{\rm Mpc}, z=0).} \tag{96}$$

where we also indicate that $\sigma_8$ is defined at the present time (in general, you can compute $\sigma^2(R)$ at any redshift you want; the result of course depends on $z$). The value of $\sigma_8$ has — also historically! — changed between about 0.6 and 1.0; today it seems to have converged around 0.8.

Note that $\sigma_8$ gives you one way to *normalize* the power spectrum: by measuring $\sigma_8$, from the distribution of galaxies for example, you can essentially determine the normalization $A$ from Eq. (163). It has been only recently, with the precision of CMB experiments, that we measure $A$ independently by studying the amplitude of fluctuations in the CMB. Measurements of $\sigma_8$ from the abundance of clusters and weak lensing, and those of $A$ from the CMB, are in good concordance. In fact, the University of Michigan cosmologists led way in using galaxy clusters to measure $\sigma_8$, getting $\sigma_8(\Omega_M/0.25)^{0.41} = 0.832 \pm 0.033$ after marginalization over all systematics (Rozo et al., ApJ 708, 645 (2010)).

**Correlation function: real space estimators.** Let us now consider find an *estimator* for the two point function — a statistical operation that we can apply to the data and extract the two point function (and, ideally, its error). Any estimator should have these desirable properties:

- The estimator should be *unbiased* — on average, it should return the correct, "true" result
- The estimator should have *minimum variance* among all choices of estimators

Finding a good estimator is sometimes more art than science. Historically, the first estimator for the correlation function $\xi(r)$ was the *Peebles-Hauser* estimator

$$\hat{\xi}_{\rm PH} = \left(\frac{N_{\rm rand}}{N_{\rm data}}\right)^2 \frac{{\rm DD}(r)}{{\rm RR}(r)} - 1 \tag{97}$$

where ${\rm DD}(r)$ is the number of pairs in the catalog in the interval $r \pm dr$, while ${\rm RR}(r)$ is the number of pairs in a *random-distribution* generated catalog in the same distance interval. The numbers $N_{\rm rand}$ and $N_{\rm data}$ are the total numbers of points (say, galaxies) in the two catalogs respectively.

Over time, estimator with better properties (smaller bias and variance) have been found. For practical purposes, it is sufficient to stick with the *Landy-Szalay* estimator

$$\boxed{\hat{\xi}_{\rm LS} = \left(\frac{N_{\rm rand}}{N_{\rm data}}\right)^2 \frac{{\rm DD}(r)}{{\rm RR}(r)} - 2\frac{N_{\rm rand}}{N_{\rm data}}\frac{{\rm DR}(r)}{{\rm RR}(r)} + 1}, \tag{98}$$

which, when $N_{\rm rand} = N_{\rm data}$, takes a more memorable form

$$\hat{\xi}_{\rm LS} = \frac{{\rm DD} - 2{\rm DR} + {\rm RR}}{{\rm RR}}. \tag{99}$$

The variance in all of these estimators, assuming we have a Poisson process, is, approximately

$$\boxed{\sigma^2_\xi(r) = \frac{1+\xi(r)}{{\rm DD}(r)} \sim \frac{1}{{\rm DD}(r)}}. \tag{100}$$

Since, however, a clustered field of e.g. galaxies is clearly not Poissonian, the actual variance can be somewhat bigger than this.

**Angular two-point correlation function.** Often in cosmology, we observe objects on the sky — the sky being a surface. In other words, we often do not have an opportunity to measure the radial distance to objects. This is of course literally true for the CMB (which "sits" on the surface of last scattering), but is also often true for the galaxy distribution, whose angular properties are more easily measured[5].

[5]To get the radial information for the galaxies, you have to measure their redshifts, but this is very time consuming if you are doing spectroscopy (measuring spectra directly), or less time consuming but more uncertain if you are using photometry (inferring redshifts from the colors).

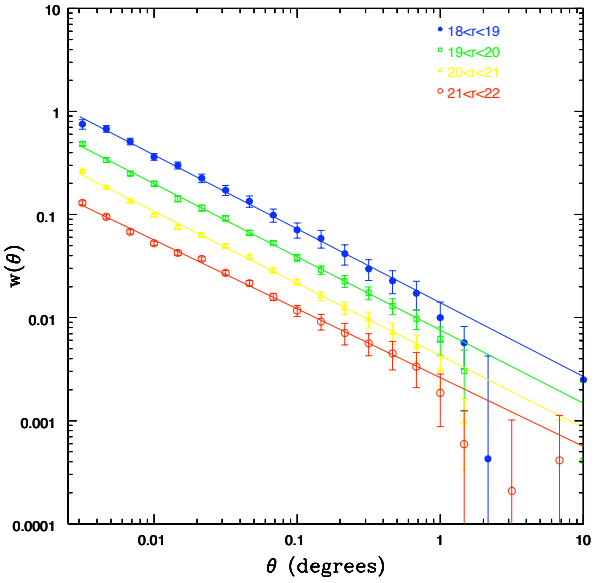


Figure 4: Angular two-point correlation function as measured in the Sloan Digital Sky Survey (Connolly et al., 2002). Notice that $w(\theta)$ depends on the magnitude of the galaxy subsample — that is, on the mean depth at which galaxies from that subsample are found.

Here we first concentrate on the angular two-point correlation function of galaxies, $w(\theta)$. In the next section we study the angular two-point function of the CMB, $C(\theta)$.

The angular two-point function is defined via

$$dP = \mathcal{N}^2(1 + w(\theta))\, d\Omega_1 d\Omega_2 \tag{101}$$

where $\mathcal{N}$ is now the mean density of points/galaxies per solid angle, and $\Omega$ is the solid angle. Clearly, $\langle N \rangle = \mathcal{N}\Omega$.

Estimators for $w(\theta)$ are simple to find. The most obvious estimator would determine the angular two point function from the map as follows

$$\hat{w}(\theta) = \frac{1}{N_{\rm pairs}(\theta)} \sum_{i,j \text{ within } \theta} \delta_i \delta_j \tag{102}$$

where we have pixelized the sky into pixels of some size (that is, some solid angle) and $\delta_i$ is the fractional overdensity in the $i$th pixel. Here the sum goes over all products of pixel overdensities that are $[\theta, \theta + d\theta]$ apart from each other, and $N_{\rm pairs}$ is the number of such pairs for each $\theta$. This formula looks a bit more complicated once we allow for the fact that pixel may be partially masked (to exclude, for example, bright sources in them), or that the selection function may be uneven (i.e. that the survey may cover different depths in different directions).

The other estimator is our old friend, which we can rewrite as

$$\hat{w}_{\rm LS}(\theta) = \frac{\rm DD - 2DR + RR}{\rm RR}. \tag{103}$$

**Mass function.** Press & Schechter (1974) stated that the likelihood for collapse of objects of a specific size or mass ($R \propto M^{1/3}$) could be computed by examining the density fluctuations

on the desired scale. They continued by using a model for the collapse of a spherical tophat overdensity to argue that collapse on scale R should occur roughly when the smoothed density on that scale exceeds a critical value $\delta_c$, of order unity, independent of $R$.

The mass within a region of size $R$ is (implicitly assuming the top-hat window function that cuts off abruptly at $R$)

$$M = \frac{4\pi}{3}\rho_M R^3 \tag{104}$$

where as usual $\rho_M \equiv \rho_M(z) = \rho_{\rm crit}\Omega_M(1+z)^3$.

Press and Schechter reasoned that, given the smoothing radius $R$, the fractional volume occupied by collapsed objects is proportional to regions whose overdensity is greater than some critical value. We have not derived this in the course, but for spherical collapse, one can show that the critical value for the collapse is

$$\boxed{\delta_c \simeq 1.686} \qquad \text{(critical overdensity for collapse)} \tag{105}$$

Therefore a quantity of interest will be fraction of collapsed objects

$$F(M) \equiv \int_{\delta_c}^{\infty} P(\delta)d\delta = \frac{1}{\sqrt{2\pi}\sigma(M)}\int_{\delta_c}^{\infty} e^{-\delta^2/(2\sigma(M)^2)}d\delta \tag{106}$$

$$\equiv \frac{1}{2}{\rm erfc}\left(\frac{\nu}{\sqrt{2}}\right) \tag{107}$$

where erfc is the *complementary error function* (see Wikipedia) and

$$\boxed{\nu \equiv \frac{\delta_c}{\sigma(M)}} \tag{108}$$

Note immediately the problem: as $\sigma(M) \to \infty$, $\nu \to 0$ and the collapsed fraction goes to 1/2, not 1. This intuitively corresponds to the fact that according to assumptions so far, only the overdensities, and not the underdensities, can lead to collapsed structures. Press and Schechter resolved this in an incredibly bold way, by multiplying the probability by a factor of 2!

The comoving number density of objects in an interval $dM$ around a mass $M$

$$\frac{dn}{d\ln M}d\ln M = \frac{\rho_{M,0}}{M}\left|\frac{dF(M)}{d\ln M}\right| d\ln M \tag{109}$$

where, since we are talking about a comoving density, we have $\rho_{M,0}$ evaluated at the present time (i.e. $\rho_M$ at an arbitrary redshift, divided by $(1+z)^3$ – gives just $\rho_{M,0}$).

After taking the derivative of Eq. (106) analytically, and include the miraculous factor of 2, we get the *Press-Schechter mass function*

$$\boxed{\frac{dn}{d\ln M} = \sqrt{\frac{2}{\pi}}\frac{\rho_{M,0}}{M}\frac{\delta_c}{\sigma}\left|\frac{d\ln\sigma}{d\ln M}\right| e^{-\delta_c^2/(2\sigma^2)}} \tag{110}$$

Note that $\sigma(M,z) = \sigma(M,0)D(z)$, so that the mass-radius conversion, from Eq. (104), only needs to be done at $z = 0$.

Note a few things

- The fact that we have a formula that quite accurately describes the abundance of halos is fantastic, since halos are inherently *nonlinear* objects (i.e. they it is NOT true that $\delta \ll 1$ at the scale of the halo).

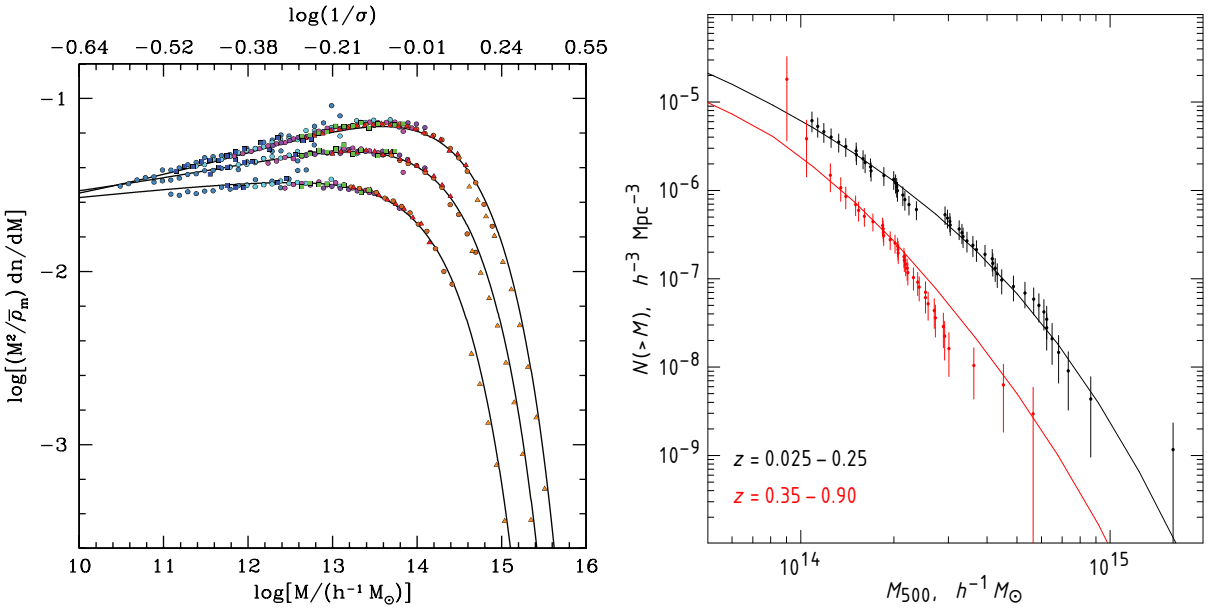


Figure 5: Left panel: "Measurements" of the mass function from an N-body simulation, together with a parameteric fit. This fit (and simulation) are accurate to better than 5%. On the y-axis, the quantity plotted is $(M/\rho_M)(dn/d\ln M)$. On the x-axis is the mass, for three definitions of 'mass' (three curves) corresponding to sum of stuff in a region out to 200, 8000 and 3200 times the mean matter density (top to bottom). Adopted from Tinker et al, ApJ 688, 709 (2008). Right panel: "Real" measurements of the mass function from the 400 square degree survey of ROSAT galaxy clusters followed up by Chandra Space Telescope. Points with error bars are data, and the lines are fits with the theoretical mass function. Adopted from Vikhlinin et al., ApJ 692, 1060 (2009).

- In particular, note that the assumption of Gaussianity was fishy, since as we know $\delta \geq -1$ by definition, and here we are talking about typical $\delta$'s of order unity. So why does the formula match the simulations so well (to a few tens of percent)? This is a subject of current research.
- The number density of objects falls exponentially with increasing mass. This is a fundamental property in our universe.
- The smoothing scale ($R$, or the corresponding mass $M$) matters. On larger scales (larger $M$), $\sigma$ is smaller, and the mass function drops sharply. There are fewer – much fewer – objects of higher mass than of lower mass.
- The PS formula is "universal", since the halo abundance depends on the cosmological model only via the rms variance $\sigma$. In other words, all of the dependence on the cosmological parameters ($\Omega_M$, $n$, $A$, $\Omega_B$, etc) is channeled through the single cosmology-dependent function $\sigma(z, M)$. There is no fundamental reason why the mass function should be universal however, and the near-universality is a hot subject of research. Recently, clear departures from universality, at $\sim 5\%$ level, have been detected by comparing theory to numerical simulations.

The determination of the mass function is done by fitting the results of N-body simulations. Huge progress in this area has been made since the work of Press and Schechter. While the PS mass function is only accurate at the $\sim 50\%$ level, the most recent fit to simulations is accurate

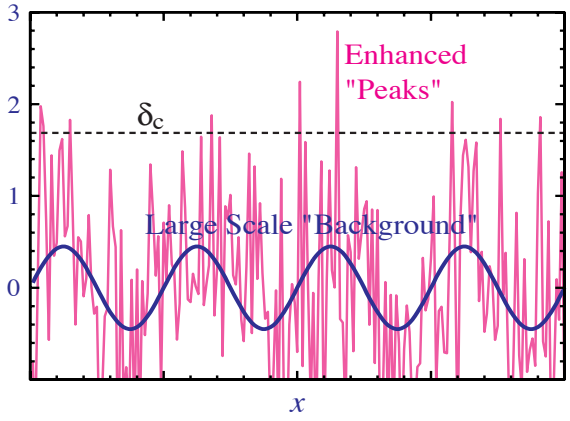


Figure 6: Figure illustrating the peak-background split. Short fluctuations (the 'peaks') live on top of long fluctuations (the 'background'), and the background raises or lowers the effective threshold for a halo to form. Adopted from Wayne Hu.

to better than 5% (Tinker et al, ApJ 688, 709 (2008); see the left panel of Fig. 5.). In this and other recent papers, clear *departures from universality* have also been detected. Upon closer scrutiny, the mass function is therefore not universal after all.

Moreover, recent results from counting galaxy clusters effectively *measure* the mass function, and agree with numerical predictions; see the right panel of Fig. 5.

**Bias of dark matter halos.** If you are standing at Mt. Everest, it is more likely that you will find another high peak near you than if you live in the Midwest of the U.S. Similarly, peaks in the density field (i.e. mostly the dark matter), which are the halos (i.e. galaxies and clusters with their DM halos) we observe on the sky, are more clustered than the the field as a whole. Schematically, the two are related with a factor $b$ that is called the bias of dark matter halos (or just bias):

$$\delta_h = b\,\delta_m \qquad \text{(definition of bias)} \tag{111}$$

Then the two corresponding power spectra are related by

$$\boxed{P_h(k,z) = b^2(k,z)P(k,z)} \tag{112}$$

where we left the possibility that bias depends on scale as well as redshift.

So what we measure in cosmology is clustering of halos (galaxies, clusters, etc); this is represented by $P_h(k)$. What we can *predict* is the clustering of dark matter, $P(k)$ (see Eq. 77 – this is where it's predicted from first principles in terms of a few fundamental parameters). The ratio between the two is the bias squared.

Can we predict the bias too, or do we have to determine it from the data concurrently with other cosmological parameters? It turns out, bias can be predicted, at least to a decent accuracy, using a clever trick called the peak-background split.

**Peak-background split.** Let us split the density fluctuation in the universe into ones of short and long wavelength. We will denote the long-wavelength perturbations as $\delta_b$ ('b' for background), and short-wavelength with $\delta_p$ ('p' for peaks):

$$\delta = \delta_b + \delta_p \tag{113}$$

We have already mentioned while discussing the Press-Schechter formalism that peaks form when the density exceeds some threshold ($\delta_c = 1.69$ in the spherical collapse model). Now, the long-wavelength fluctuations form a background on which the peaks form, and therefore $\delta_b$ serves the role of changing the threshold from $\delta_c$ to $\delta_c - \delta_b$. See Fig. 6.

Now let us use the power of the peak-background split, and expand the number density in Taylor series, assuming the Press-Schechter mass function, $n(\nu) \propto \nu \exp(-\nu^2/2)$:

$$\begin{aligned} n(\nu + \delta\nu) = n\left(\frac{\delta_c - \delta_b}{\sigma}\right) &\approx n(\nu) + \frac{dn}{d\nu}\frac{d\nu}{d\delta}(-\delta_b) \\ &= n(\nu) + \left(\frac{1}{\nu} - \nu\right) n(\nu) \left(-\frac{\delta_b}{\sigma}\right) \\ &= n(\nu)\left[1 + \frac{\nu^2 - 1}{\nu\sigma}\,\delta_b\right] \end{aligned} \tag{114}$$

Therefore, $\delta n/n = (\nu^2 - 1)/(\nu\sigma)\,\delta_b \equiv b_L \delta_b$. The quantity $b_L$ is the *Lagrangian bias*, which is bias in coordinates moving with the expansion. We are ultimately interested in the *Eulerian bias* $b_E$, which is related to the Lagrangian bias via (e.g. Mo & White, 1996)

$$b_E (\equiv b) = b_L + 1 \tag{115}$$

Therefore,

$$\boxed{b(M) \simeq 1 + \frac{\nu^2 - 1}{\delta_c}} \qquad \text{(bias from peak} - \text{background)} \tag{116}$$

where, again, remember that $\nu(M) = \delta_c/\sigma(M)$. This expression, which is in good agreement with bias derived from both N-body simulations and observations, shows that the *bias for more massive objects is bigger*.

# Statistical Methods in Cosmology and Astrophysics

**Dragan Huterer, University of Michigan.**

**Recommended reading:** There are some excellent resources on topics covered here.

- "Statistics in theory and practice", Robert Lupton — *A compact, yet detailed book, indispensable for serious statistics practitioners.*
- "Numerical Recipes - the Art of Scientific Computing", Press, Teukolsky, Vetterling & Flannery — *Famous magnum opus that explains numerical and statistical topics very clearly. Also comes with computer code that you can download and use in your research.*
- "Bayes in the sky: Bayesian inference and model selection in cosmology", R. Trotta, arXiv:0803.4089 — *A fairly complete scripture of various Bayesian techniques from one of the Apostles.*
- "A practical guide to Basic Statistical Techniques for Data Analysis in Cosmology", L. Verde, arXiv:0712.3028, and "Statistical methods in Cosmology", arXiv:0911:3105 by same author — *Good, broad overviews of various numerical/statistical topics in cosmology with a good list of references.*
- "Karhunen-Loeve Eigenvalue Problems in Cosmology: How Should We Tackle Large Data Sets?", M. Tegmark, A. Taylor and A. Heavens, ApJ, 480, 22 (1997) — *one of the papers that introduced the Fisher matrix to cosmology; explained well and major bonus materials on data compression in cosmology if you are interested.*
- "Unified approach to the classical statistical analysis of small signals", G.J. Feldman and R.D. Cousins, PRD, 57, 3873 (1998) — *If you would like apply a frequentist approach to a problem, read this very clear and important paper, which also gives applications to neutrino oscillation data.*

**Random variables.** A random variable (or stochastic variable) $X$ is a variable whose value is subject to variations due to chance. It is a quantity whose value is not fixed, but which can take on different values whose likelihood is described by a probability distribution.

A random variable's possible values might represent the possible outcomes of a yet-to-be-performed experiment, for example.

**Likelihood.** I first note I will use the terms probability and likelihood somewhat interchangeably here. I don't think there is a difference between the two.

The following fundamental rules of probability hold:

- $P(X) \geq 0$.
- $\int_{-\infty}^{\infty} P(X) dX = 1$
- $P(X_2) = \int P(X_2|X_1) P(X_1) dX_1$

Let us comment a bit more on the last one of these rules. Here $P(X_2|X_1)$ means probability of even $X_2$ given event $X_1$. If the events are independent, then $P(X_2|X_1) = P(X_2)$ and the third item above becomes a tautology, $P(X_2) = P(X_2)$.

Finally, note that one can write, in full generality

$$P(X_1, X_2) = P(X_2|X_1)P(X_1) = P(X_1|X_2)P(X_2), \tag{117}$$

where here we have "probability of $X_1$ and $X_2$".

The distribution is often described by its first moment, the *mean* of the distribution, which is defined as:

$$\mu \equiv \bar{x} \equiv \langle X \rangle = \int_{-\infty}^{\infty} XP(X)\, dX \quad \text{(mean)} \tag{118}$$

Related quantities are the median and the mode:

$$\frac{1}{2} = \int_{-\infty}^{X_{\rm median}} P(X)\, dX \quad \text{(median)} \tag{119}$$

$$\left.\frac{dP}{dX}\right|_{\rm x_{mode}} = 0 \qquad \text{(mode)}. \tag{120}$$

Finally, it is useful to see how you can *marginalize* (integrate) over some variables to be left with the marginalized likelihood in others. Imagine you have two variables $X$ and $Y$ with the joint probability function $P(X, Y)$, and you would like to find out what's the probability in $X$ alone. Easy:

$$P(X) = \int P(X, Y) dY \qquad \text{(marginalization)}. \tag{121}$$

**Variance and higher moments.** Variance is a natural measure of the width (squared) of a distribution. The variance of the random variable $X$ that is described by the distribution $P(X)$ is given by

$$\mathrm{Var}(X) \equiv \sigma^2 \equiv \langle (X - \mu)^2 \rangle = \int_{-\infty}^{\infty} (X - \mu)^2 P(X)\, dX \quad \text{(variance)} \tag{122}$$

where $\mu$ is the mean. In other words, variance is the second moment of the distribution around the mean. Note that the n-th moment of the distribution is defined by $\langle X^n \rangle \equiv \int X^n P(X) dX$.

Similarly to variance, one can define higher moments of the distribution *around the mean*. For example skewness

$$S \equiv \left\langle \left( \frac{X - \mu}{\sigma} \right)^3 \right\rangle \qquad \text{(skewness)} \tag{123}$$

measures the asymmetry of the probability distribution around the mean, while kurtosis

$$K \equiv \left\langle \left( \frac{X - \mu}{\sigma} \right)^4 \right\rangle \qquad \text{(kurtosis)} \tag{124}$$

measures the "peakedness" of $P$.

**Estimators.** In statistics, estimator is a procedure for calculating the desired statistical quantity (e.g. mean and variance of some random variable; expected salary; age of the universe; etc) based on observational data.

A single quantity can be estimated using many different ways, i.e. using different estimators. Typically we strive for an estimator with minimal bias (expected deviation from the truth) and minimal variance (reported error bar).

The simplest example is an estimator of the mean of some random variable $X$, given measurements $x_1, x_2, \ldots, x_N$. The obvious (and good) estimator of the mean is

$$\hat{\mu} = \frac{\sum_{i=1}^{N} x_i}{N} \tag{125}$$

where the hat stands for the "estimate".

OK, now how do you estimate the variance? You can try the naive estimator $\widehat{\mathrm{Var}} = \sum(x_i - \hat{\mu})^2/N$, but this estimator turns out to be *biased* because the measurements $x_i$ and the mean estimate $\hat{\mu}$ are correlated; see Lupton section 5. Instead, an *unbiased* estimator of the variance is

$$\widehat{\mathrm{Var}} = \frac{\sum_{i=1}^{N}(x_i - \hat{\mu})^2}{N-1}. \tag{126}$$

In general, if you can find the *best unbiased estimator* for the quantity of your interest (zero bias, smallest variance of the estimate) then you are doing really well.

**Gaussian (or, normal) distribution.** Gaussian distribution is by far the most standard statistical distribution, is a gold standard in statistics, and is commonly found in phenomena describing physics, astronomy, and most other natural and social sciences. Gaussian distribution is also the simplest distribution to work with.

The probability density function (PDF) of the Gaussian distribution in one dimension (one parameter) is

$$P(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right] \tag{127}$$

where $\mu$ is the mean of the distribution, and $\sigma$ is the standard deviation. The variance is $\sigma^2$; skewness is zero, and kurtosis (and all even higher moments) can be easily expressed in terms of the variance. For multiple parameters, ordered in a vector $\mathbf{x}$ with mean $\mu$ and covariance matrix $C \equiv \langle \mathbf{x}\mathbf{x}^T \rangle - \langle \mathbf{x} \rangle \langle \mathbf{x}^T \rangle = \langle (\mathbf{x} - \mu)(\mathbf{x} - \mu)^T \rangle$, the Gaussian PDF generalizes to

$$P(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} |\det C|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mu)^T C^{-1} (\mathbf{x} - \mu)\right]. \tag{128}$$

where, recall, $\mu = \langle \mathbf{x} \rangle$ is the mean (vector).

**Poisson Distribution.** Poisson (pronounced `pwason`, not `poi'son`) distribution expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate and independently of the time since the last event. An example of the Poisson distribution is the number of cars passing on a street in a given time of the day; the expected number per time interval is fixed, but the actual number obviously fluctuates (i.e. is stochastic).

The Poisson PDF is given by

$$P(X, n) = \frac{n^X e^{-n}}{X!} \tag{129}$$

where $n$ is the expected number of events, and $X$ is the observed number. The first two moments are

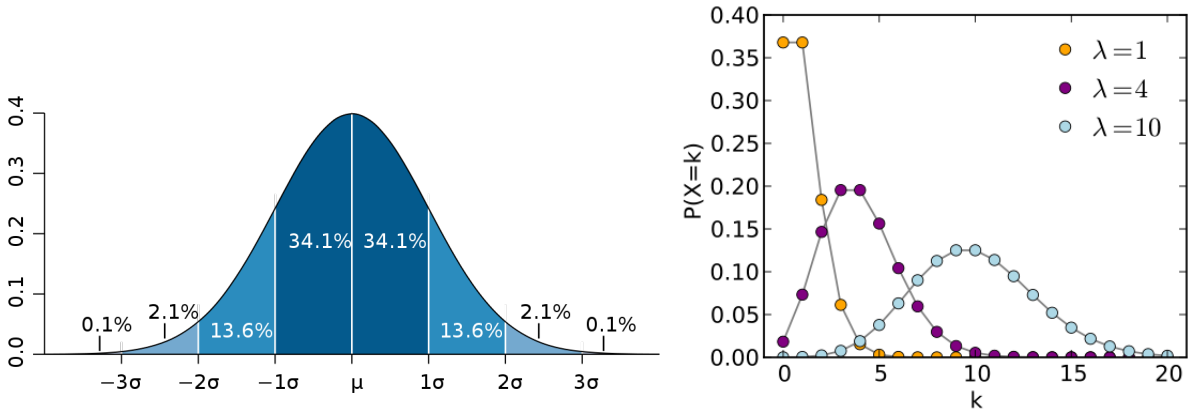


Figure 7: Left panel: Gaussian distribution. Right panel: Poisson distribution, for a few different values of mean expected counts. Both plots are adopted from Wikipedia.

- mean: $n$
- variance $n$
- When $n \gg 1$, Poisson distribution starts looking a lot like the Gaussian distribution.

Poisson distribution is extremely common in cosmology and astrophysics, especially when you count things (photons in a detector, galaxy clusters on the sky, etc).

**Chi-square distribution.** Chi-square (or chi-squared, or $\chi^2$) distribution is one that a *sum of squares of gaussian variables have*. That is, if

$$Y = X_1^2 + X_2^2 + \ldots + X_n^2 \tag{130}$$

where $x_i$ are Gaussian random variables with mean zero and variance of one, then $y$ has chi-square distribution with $n$ degrees of freedom:

$$P(Y) = \frac{1}{2^{n/2}\Gamma(n/2)} Y^{n/2-1} e^{-Y/2}. \tag{131}$$

This distribution has the following lowest moments:

- mean: $n$
- variance $2n$
- When $n \gg 1$, chi-square distribution starts looking a lot like the Gaussian (with mean and variance as above, of course).

In applications in cosmology, we often have likelihoods that are a function of

$$\mathcal{L} \propto \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{X_i - \mu_i}{\sigma_i}\right)^2\right] \equiv \exp\left[-\chi^2/2\right] \tag{132}$$

**Central Limit Theorem.** Let $X_1$, $X_2$, $X_3$, ..., $X_n$ be a sequence of $n$ independent and identically distributed (iid) random variables each having finite values of expectation $\mu$ and

variance $\sigma^2 > 0$. The central limit theorem states that as the sample size n increases the distribution of the *arithmetic mean* of these random variables approaches the normal distribution with a mean $\mu$ and variance $\sigma^2/n$ irrespective of the shape of the common distribution of the individual terms $X_i$.

The CLT is very useful in astrophysics and cosmology, since it often simplifies the analysis. Examples are:

- a collection of data points whose joint distribution can be considered Gaussian even though individual points are definitely not Gaussian distributed around their means;
- looking at density fluctuations in the distribution of galaxies (or CMB temperature) in a patch of the sky; even though the individual modes are not Gaussian distributed around zero, the overall distribution is accurately described by the Gaussian.

**Chi squared.** Let is develop the first example from above a little more. Consider measurements of a number of quantities (sometimes people talk about "observables") $X_i$, where $i = 1, 2, \ldots, n$ is the number of these observables. These observables may not be Gaussian-distributed random variables. Yet if $n$ is large, it turns out that their joint distribution is well described by a Gaussian (in this observable space). Defining

$$\chi^2 = \sum_{i=1}^{n} \left( \frac{X_i - \mu_i}{\sigma_i} \right)^2 \tag{133}$$

the likelihood is well described by $\mathcal{L} \propto \exp(-\chi^2/2)$ in the limit of large $n$.

Let us now say that these observable inform us about some theoretical parameters (say $\Omega_M$ or $H_0$ or electron mass or whatever). Let $\mathbf{p}$ be the vector of these theoretical parameters. Then, if the likelihood is Gaussian in the observables $X_i$, it does *not* necessarily follow that it is Gaussian in the parameters $p_j$!

Make sure you understand the difference between the two spaces of random variables:

1. **Data** (measured quantities): $\mathbf{X} = \{X_i\}$. By the CLT, it is usually OK to assume a Gaussian joint distribution in the data, $\mathcal{L}(\mathbf{X})$, even if individual data points are not Gaussian-distributed.
2. **Model parameters** (theoretical parameters): $\mathbf{p} = \{p_j\}$. It is usually *not* OK to assume that $\mathcal{L}(\mathbf{p})$ is Gaussian (due to e.g. nonlinear relations between the true underlying observable quantities and theory parameters).

Generally, it makes sense to write

$$\mathcal{L} = \mathcal{L}(\bar{\mathbf{X}}(\mathbf{p})) \tag{134}$$

where $\bar{\mathbf{X}}$ are the true, underlying observable quantities: the latter inform us about the theoretical parameters.

**Maximum likelihood, minimum $\chi^2$.** Once the best fit parameters are obtained, how can one represent the confidence limit or confidence region around the best fit parameters? A reasonable choice is to find a region in the $n$-dimensional parameter space (remember that $n$ is the number of parameters) that contain a given percentage of the probability distribution. In most cases one wants a compact region around the best fit vales.

The simplest logic is to use the *likelihood ratio*. The likelihood at a particular point in parameter space is compared with that at the best fit value, $\mathcal{L}_{max}$ where likelihood is maximized.

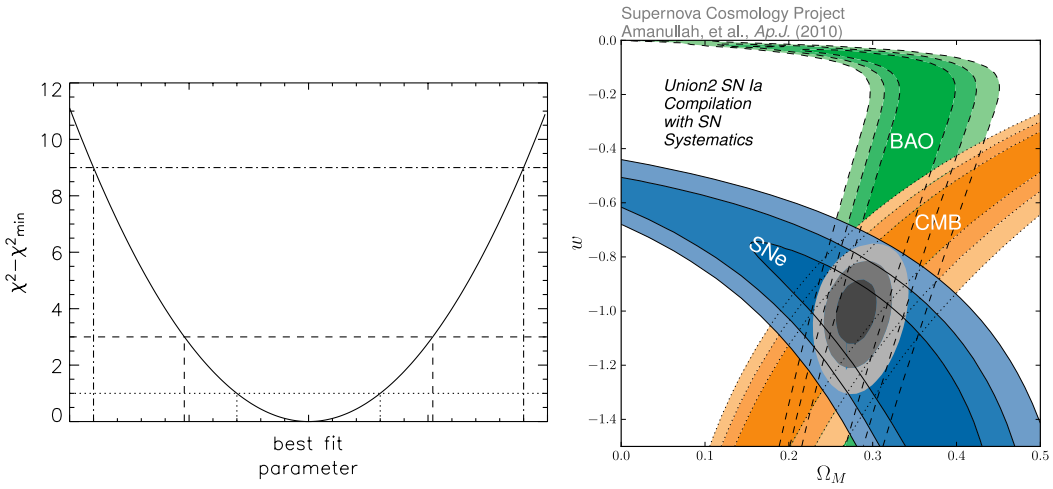


Figure 8: Left panel: example of a one-dimensional chisquare for a Gaussian distribution as a function of a parameter and corresponding 68.3%, 95.4% and 99.7% confidence levels. Adopted from Verde, arXiv:0911.3105. *Right panel*: Constraints upon $\Omega_{\rm M}$ and constant $w$ in the fiducial dark energy model using the same data sets. From Supernova Cosmology Project's Union2 compilation of 557 SN (Amanullah et al., 2010). Note that in a practical application even if the data have gaussian errors the errors on the parameter may not be well described by multi-variate Gaussians (thus the confidence regions are not ellipses).

Thus those parameter values are favored that lead to the likelihood ratio $[\mathcal{L}(\mathbf{p})/\mathcal{L}_{max}]$ above a given threshold. The threshold can be chosen so that the enclosed total probability is, say, 68.3% of the total (see below).

Note that, for a purely Gaussian distribution in the parameters (which we often assume), $L \propto \exp(-\chi^2/2)$, and maximum likelihood $\mathcal{L}_{max}$ corresponds to minimum chi square, $\chi^2_{\rm min}$. Then the natural choice is given by regions of constant $\chi^2$ boundaries

$$\chi^2 - \chi^2_{\rm min} = -2\ln\left[\frac{\mathcal{L}(\mathbf{p})}{\mathcal{L}_{\rm max}}\right] \tag{135}$$

Note that there may be cases (when the $\chi^2$ has more than one minimum) where the confidence regions are disjoint and/or of weird shapes. For multi-variate Gaussian distributions in the theory parameters, however, the confidence regions are always ellipsoids (e.g. ellipses in $N_{\rm params} = 2$ dimensional parameter space). Note that the fact that the data have Gaussian errors does *not* imply that the parameters will have a Gaussian probability distribution.

There is a subtlety to point out here. In cosmology the data may be Gaussian-distributed and still the $\chi^2$ and likelihood ratio analysis may give different results. This happens because in identifying likelihood and chisquare we have neglected the term $[(2\pi)^{n/2}|\det C|^{1/2}]^{-1}$. If the covariance does not depend on the model or model parameters, this is just a normalization factor which drops out in the likelihood ratio. However in cosmology the covariance often depends on the model: this happens for example if your random variable $X$ is the overdensity $\delta \equiv \delta\rho/\rho$, then its mean is zero (and clearly doesn't depend on the cosmological parameters), while the covariance of $\delta$ is the two-point correlation function $\xi(r)$ in real space (or power spectrum $P(k)$ if you are talking about $\delta = \delta_{\mathbf{k}}$ in Fourier space), and the latter quantity depends on the cosmological parameters.

**Likelihood confidence levels – a recipe.** Our goal is to specify how to define the com-

monly used 68.3%, 95.4% and 99.7% confidence levels for some parameter. [Exactly the same procedure holds for the CL for two or more parameters jointly – it's just applied to the likelihood in 2D, 3D etc – but for simplicity we keep the discussion for one parameter.] Note that, in the community, these limits are still called 1-$\sigma$, 2-$\sigma$ and 3-$\sigma$ ranges even though they really are 68.3%, 95.4% and 99.7% respectively. Of course, the two nomenclatures agree for a Gaussian likelihood in the parameters, but most often the likelihood is not gaussian. So just remember that "sigmas" is just a lazy way of quoting the percentages that *would* obtain for a Gaussian.

The general prescription to compute the confidence levels is as follows:

1. Find the best-fit parameter value by finding the maximum-likelihood. Call this minimum value $\mathcal{L}_{\rm max}$

2. Going to values smaller and larger than this value, go "down the likelihood" until you enclose 68.3% of the total – that is, find $a$ and $b$ so that $\mathcal{L}(a) = \mathcal{L}(b)$ and

$$\int_a^b \mathcal{L}(x)dx = 0.683 \int_{-\infty}^{\infty} \mathcal{L}(x)dx \tag{136}$$

   where of course the full range on the rhs may be different in a particular situation (e.g. if $x$ is some mass, it starts at zero value).

3. If desired, repeat for another confidence level (say, 95.4%, and find $c$ and $d$ such that

$$\int_c^d \mathcal{L}(x)dx = 0.954 \int_{-\infty}^{\infty} \mathcal{L}(x)dx \tag{137}$$

Then you would say, for example, that "2-$\sigma$ allowed range for $x$ is $[c, d]$".

Note that this easily generalizes in several directions. For multiple parameters, you walk in the full-parameter space away from maximum likelihood (so e.g. for two parameters you have a region that encloses these same fractions of the total area under the likelihood). And if your likelihood does not go to zero sufficiently fast at one end, e.g. the low end of $x$, then you will only have the upper limit on $x$, but the procedure is precisely the same above.

**Goodness-of-fit.** Completely separately from finding the parameter values (and their e.g. 68% ranges) of a model given some data, there is a basic question whether the model itself is a good fit to the data. In fact, when you are fitting some model to the data (e.g. standard cosmological parameters to the DES galaxy clustering and weak lensing data), you have to perform two separate calculations:

1. Find the best-fit values of the parameters, as well as their errors/ranges (or really, their full parameter covariance); and

2. Is the model, evaluated at the best-fit parameter values, a good fit to the data? Yes or no.

We now discuss the latter of these questions.

If the model is a good fit to the data, we expect the data to scatter around the model symmetrically and "on average" by 1-sigma (where sigma is the error on each measurement). More generally, this is codified by the simple requirement that

$$\boxed{\chi^2_{N_{\rm dof}} \simeq N_{\rm dof}} \qquad \text{(when model is a good fit to data)} \tag{138}$$

where $\chi^2_{N_{\rm dof}}$ is something like $(\mathbf{x}-\mathbf{x}_{\rm model})^T C^{-1}(\mathbf{x}-\mathbf{x}_{\rm model})$, while $N_{\rm dof}$ is the number of degrees of freedom:

$$N_{\rm dof} \equiv N_{\text{data points}} - N_{\text{fitted parameters}}. \tag{139}$$

You can use the properties of the chi-squared distribution — particularly the fact that $\mathrm{Var}(\chi^2_{N_{\rm dof}}) = 2N_{\rm dof}$, to quantify when the fit is not good. For example, DES Y1 key paper (`https://arxiv.org/abs/1708.01530`) contained 457 measurements, and 26 free parameters (6 cosmological parameters and 20 nuisance ones), leading to $N_{\rm dof} - 457 - 26 = 431$. Therefore, you expect chi squared in the range of $\chi^2 \simeq 431 \pm \sqrt{2\times 431} \simeq 431 \pm 30$. Anything *much* more than that, and either the model is not a good fit, or you have residual systematic errors, or else you underestimated the error bars, or... something went wrong.

A lot of the time $\chi^2$ is higher than that expected due to any of the aforementioned reasons. Very rarely, it also happens that chi squared is *lower* than that expected, and almost always this means that you *over*-estimated the error bars.

**Bayesian vs. frequentist.** There are two principal approaches to statistics, and their competition is as famous as that between the Montagues and Capulets, the Hatfields and McCoys, or the Lakers and the Celtics. These are the Bayesian and frequentist approaches.

Frequentist interpretation of probability defines an event's probability as the limit of its relative frequency in a large number of trials. So I observe the event unfold many times and, in the limit when that number goes to infinity, the relative frequency of its outcome becomes its probability.

Bayesian probability interprets the concept of probability as 'a measure of a state of knowledge, and not as a frequency. One of the crucial features of the Bayesian view is that a probability can be assigned to a hypothesis, which is not possible under the frequentist view, where a hypothesis can only be rejected or not rejected.

In short, the difference is

- Bayesian: data are fixed, model is repeatable.
- Frequentist: model is fixed, data are repeatable.

**Bayesian statistics.** More formally, the Bayesian probability calculus makes use of Bayes' formula - a theorem that is valid in all common interpretations of probability - in the following way:

$$\boxed{P(M|D) = \frac{P(D|M)\,P(M)}{P(D)}} \quad \text{(Bayes}' \text{ theorem)} \tag{140}$$

where $M$ represents a model (or a hypothesis) made up of the parameter vector $\mathbf{p}$, and $D$ is data. Here

- $P(M)$ is a **prior** probability of $M$ the probability that $M$ is correct before the data $D$ was seen
- $P(D|M)$ is the conditional probability of seeing the data D given that the model $M$ is true. $P(D|M)$ is actually the **likelihood** (of the data, given the model).
- $P(D)$ is the *a priori* probability of witnessing the data $D$ under all possible models. It is a normalizing constant that only depends on the data, and which in most cases does not need to be computed explicitly. This quantity however plays an important role in model

selection; is called the *Bayesian evidence*. It is given by the likelihood integrated (summed) over all model-parameter values:

$$P(D) = \int P(D|\mathbf{p}, M) P(\mathbf{p}|M)\, d\mathbf{p} \qquad \text{(Bayesian evidence)} \tag{141}$$

$$\equiv \int \mathcal{L}(\mathbf{p})\, \text{Prior}(\mathbf{p})\, d\mathbf{p} \tag{142}$$

where in the second line we simplified the notation a little. Note that the evidence may be difficult to numerically evaluate, since it integrates the likelihood over the often huge, multi-dimensional parameter space.

- $P(M|D)$ is the **posterior** probability: the probability that the model is true, given the data and the previous state of belief about the models.

The key thing to note is that we are most often interested in the probability of a model given data, $P(M|D)$, while what we can most often calculate from the data is the likelihood of the data given the model, $P(D|M)$. Bayes' theorem lets you go form from the latter to the former. Note that the two are very similar if the data are very *informative*, so that the nature of the prior in the model space doesn't matter much. However, when the data are not very informative, choice of the prior may play a role.

Bayesian approach has many advantages, and has been near-universally accepted in cosmology since the data boom in the 1990s.

- Bayesian approach allows easy incorporation of different data sets. For example, you can have one data set impose an effective prior on the model space $M$, and then this prior probability is updated with a new data set using the Bayes' theorem.
- In frequentist statistics, a model can only be rejected or not rejected. In Bayesian statistics, a probability can be assigned to a model (provided you know or can calculate the marginal probability of the data, $P(D)$).

**Bayesian-frequentist example.** Say for example that we have a measurement of the Hubble constant of $(72 \pm 8)$km/s/Mpc. What would the Bayesian and the frequentist say?

- Bayesian: the posterior distribution for $H_0$ has 68% if its integral between 64 and 80km/s/Mpc. The posterior can be used as a prior on a new application of Bayes' theorem.
- Frequentist: Performing the same procedure will cover the real value of $H_0$ within the limits 68% of the time. But how do I repeat the same procedure (generate a new $H_0$ out of the underlying model) if I only have one Universe?

Let us give another example. Say I would like to measure $\Omega_M$ and $\Omega_\Lambda$ from SN data (let us ignore $\mathcal{M}$ for the moment and just assume these two parameters). What would the two statisticians do?

- Bayesian: Take some prior (say, uniform prior in both $\Omega_M$ and $\Omega_\Lambda$). Then, for each model $M = (\Omega_M, \Omega_\Lambda)$, compute the likelihood of the data, $P(D|M)$ using, for example, the chi-square statistic. Obtain the posterior probability on the two parameters using Bayes' theorem; $P(M|D) \propto P(D|M)\, P(M)$.

- Frequentist: Calibrate your statistic by assuming a model within the range you are exploring (say, $\Omega_M = \Omega_\Lambda = 0.5$) and running many realizations of SN data with that underlying model. Each realization $i$ of the data (points, and errors) will given you $\chi^2_i$. Histogramming $\chi^2_i$ will calibrate the likelihood. Now calculate the $\chi^2$ statistic for the *real* data, assuming the same model, and compare to the histogram — this will give you a (relative likelihood) for that model. Repeat for each model $M = (\Omega_M, \Omega_\Lambda)$.

The latter approach is also called the Feldman-Cousins approach, referring to an excellent paper that I encourage you to read (see the references at the beginning of this section). It is computationally very demanding, since it requires a suite of realizations of data for *each* model $M$. To make it less demanding, you can hope for the best and assume the histogram of the statistic to be the same for each model, and only do it for one model.

**So what prior do I use?** In general, the results will depend on the prior. For example, you can consider using a flat prior on some model parameter $p$ (equal probability per $dp$), or a prior flat in the log of $p$ (so equal probability per $d\ln p$). However, *when* the data is very informative, it will completely dominate over the prior and the prior itself will be irrelevant. In other words, when you see people arguing about which prior is "more physical" given that they lead to different final parameter constraints, you conclude that the data they are using is quite weak and probably cannot lead to robust cosmological parameter constraints.

**Bayesian model comparison.** The Bayesian approach enables easy hypothesis testing, and in particular comparisons of different models, for example answering questions like

- Given some data, what is the preference for models with dark energy ($\Omega_{\rm DE} > 0$) compared to those without ($\Omega_{\rm DE} = 0$)?
- Given some data, what is the likelihood that the age of the universe is greater than 10 Gyr?

et cetera.

Let us consider two models, $M_0$ and $M_1$, that we would like to compare. We are really interested in the ratio of the posterior probabilities, or *posterior odds*, given by

$$\frac{P(M_0|D)}{P(M_1|D)} = \frac{P(D|M_0)}{P(D|M_1)}\frac{P(M_0)}{P(M_1)} \equiv B_{01}\frac{P(M_0)}{P(M_1)} \tag{143}$$

and the *Bayes factor* $B_{01}$ is the ratio

$$B_{01} \equiv \frac{P(D|M_0)}{P(D|M_1)} \quad \text{(Bayes factor)} \tag{144}$$

$$\equiv \frac{\int P(D|\mathbf{p}_0, M_0)P(\mathbf{p}_0|M_0)\,d\mathbf{p}_0}{\int P(D|\mathbf{p}_1, M_1)P(\mathbf{p}_1|M_1)\,d\mathbf{p}_1}. \tag{145}$$

Here $\mathbf{p}_0$ and $\mathbf{p}_1$ are the parameters making up models $M_0$ and $M_1$, respectively, and $P(\mathbf{p}_0|M_0)$ and $P(\mathbf{p}_1|M_1)$ are *their* prior distributions.

A value $B_{01} > (<)\ 1$ represents an increase (decrease) of the support in favour of model 0 versus model 1 given the observed data. From Eq. (143) it follows that the Bayes factor gives the factor by which the relative odds between the two models have changed after the arrival of the data, regardless of what we thought of the relative plausibility of the models before the data, given by the ratio of the prior models' probabilities. Therefore "the relevant quantity to update our state of belief in two competing models is the Bayes factor" (Trotta).

| $\lvert \ln B_{01} \rvert$ | Odds | Probability | Strength of evidence |
|---|---|---|---|
| $< 1.0$ | $\lesssim 3:1$ | $< 0.750$ | Inconclusive |
| 1.0 | $\sim 3:1$ | 0.750 | Weak evidence |
| 2.5 | $\sim 12:1$ | 0.923 | Moderate evidence |
| 5.0 | $\sim 150:1$ | 0.993 | Strong evidence |

Table 1: Empirical scale for evaluating the strength of evidence when comparing two models, $M_0$ versus $M_1$ (so–called "Jeffreys' scale") for the Bayes factor. Threshold values are empirically set, and they occur for values of the logarithm of the Bayes factor of $|\ln B_{01}| = 1.0$, 2.5 and 5.0. The right–most column gives our convention for denoting the different levels of evidence above these thresholds. The probability column refers to the posterior probability of the favoured model, assuming non–committal priors on the two competing models, i.e. $P(M_0) = P(M_1) = 1/2$ and that the two models exhaust the model space, $P(M_0|D) + P(M_1|D) = 1$. [Adopted from Trotta, arXiv.0803.4089].

Bayes factors are usually interpreted against the Jeffreys' scale for the strength of evidence, given in Table 1. This is an empirically calibrated scale, with thresholds at values of the odds of about $3:1$, $12:1$ and $150:1$, representing weak, moderate and strong evidence, respectively. A useful way of thinking of the Jeffreys' scale is in terms of betting odds — many of us would feel that odds of $150:1$ are a fairly strong disincentive towards betting a large sum of money on the outcome. Also notice from Table 1 that the relevant quantity in the scale is the logarithm of the Bayes factor, which tells us that evidence only accumulates slowly and that indeed moving up a level in the evidence strength scale requires about an order of magnitude more support than the level before.

Bayesian model comparison *does not* replace the parameter inference step (which is performed within each of the models separately). Instead, model comparison extends the assessment of hypotheses in the light of the available data to the space of theoretical models, as evident from Eq. (143)

**Markov chain Monte Carlo.** Say you want to constrain $N$ cosmological parameters; let us take $N = 10$ typical for cosmology. Say, for simplicity, that you want to allow each parameter to take $M$ discrete values; let us take $M = 10$ which is the bare barest minimum you would want to do. Then the total number of models to explore (and calculate observables for) is $M^N = 10^{10}$, which is huge — this might be doable for a simpler data set, but if you consider running CAMB (which actually only takes seconds per model), this is about 100 years. And if you want to allow the still-modest $M = 20$ values per parameter, then likelihood calculations would take 100,000 years, which means that an early Neanderthal starting the chains would make it just in time for his paper to be published this year.

Markov chain Monte Carlo (MCMC) methods are an incredibly powerful tool to overcome these problems[6]. MCMC are a class of algorithms for sampling from probability distributions based on constructing a Markov chain that has the desired distribution as its equilibrium distribution. The state of the chain after a large number of steps is then used as a sample from the desired distribution. The quality of the sample improves as a function of the number of steps.

Instead of going exponentially with the number of parameters, the MCMC calculation goes approximately linearly with $N$.

[6] A *Markov process* (or a Markov chain) is a process where the future states only depend on the present state, but not on the past states.

Usually it is not hard to construct a Markov Chain with the desired properties. The more difficult problem is to determine how many steps are needed to converge to the stationary distribution within an acceptable error. A good chain will have rapid mixing - the stationary distribution is reached quickly starting from an arbitrary position.

**MCMC: the science.** We will only consider the *Metropolis-Hastings* algorithm here, which is the most simple variant of MCMC. Those if you interested in parameter analyses are advised to take a look at the Gibbs sampler as well (for example) and read about it.

The Metropolis-Hastings algorithm draws samples from the likelihood (or probability distribution) $\mathcal{L}(\mathbf{x})$. How does it do that? The algorithm generates a Markov chain where each state $\mathbf{x}^{t+1}$ depends only on the previous state $\mathbf{x}^t$. The algorithm uses a proposal density $Q(\mathbf{x}'|\mathbf{x}^t)$ which depends on the current state $x^t$ to generate a new proposed sample $x'$. This proposal is either accepted or rejected as the next value (so, $\mathbf{x}^{t+1} = \mathbf{x}'$) or rejected (so, $\mathbf{x}^{t+1} = \mathbf{x}^t$) according to the following rule:

- calculate the ratio of the likelihood distribution at the proposed point to that at the current point
$$r \equiv \frac{\mathcal{L}(\mathbf{x}')}{\mathcal{L}(\mathbf{x}^t)} \tag{146}$$
- If $r > 1$ (that is, if the proposed point's likelihood is greater than the current), move to the new point: $\mathbf{x}^{t+1} = \mathbf{x}'$
- If $r < 1$ (that is, if the proposed point's likelihood is smaller than the current), then draw a random number $\alpha$ from a uniform distribution $U[0, 1]$.
  - If $\alpha < r$, move to the proposed point; $\mathbf{x}^{t+1} = \mathbf{x}'$.
  - If $\alpha > r$, do *not* move to the proposed point; $\mathbf{x}^{t+1} = \mathbf{x}^t$.

One can mathematically prove that the algorithm based on this rule converges to the desired *true* likelihood distribution $\mathcal{L}(\mathbf{x})$.

So the step-by-step instructions for the Metropolis-Hastings algorithm are as follows:

1. Select some proposal function $Q(\mathbf{x}'|\mathbf{x}^t)$ that tells you how to step in $\mathbf{x}$; for example this could be a multi-variate Gaussian with widths in parameters comparable to the *guessed* errors in parameters $\mathbf{x}$.
2. Start at some (randomly chosen) point in parameter space $\mathbf{x}^{t=0}$.
3. Make a proposal step, so you are considering some $\mathbf{x}'$.
4. Evaluate the proposed likelihood $\mathcal{L}(\mathbf{x}')$ and follow the MH algorithm above to decide if you are moving to $\mathbf{x}'$ or not.
5. Goto step 3, repeat.
6. End when you obtain convergence according to a criterion that you can impose.

It turns out that the 'weight' at each point in parameter space – the number of times you waited and did not move (e.g. if you got somewhere and then had reject-reject-accept, then weight=3) – is proportional to the true posterior likelihood that you would like to recover. Typically, the number of steps you need for convergence is something like 100,000 or a million

(remember, it scales linearly with $N_{\rm params}$), which is very small compared to the exponential number going as $10^{N_{\rm params}}$ that you'd need with a brute-force grid-type exploration of the parameter space.

**MCMC: the art.** The Metropolis-Hastings algorithm (above) is about 10 lines of computer code. So what's all the fuss about? Well, to make a successful MCMC, you need to take care of a number of things.

1. You need to assure that the *burn-in* stage is not included in the final results. This typically means running the MCMC for a number of steps (say, 10,000), discarding those results, and then doing a "production run" (with, say, a million steps)

2. The parameter excursions should actually not be deterministic, but equal to estimated 1-$\sigma$ errors (or square roots of eigenvalues) *times* the Gaussian normal variable of mean zero and variance one. That is, the excursion in $i$-th parameter at step $t$ is

$$\text{step}_i^{(t)} = \sigma_i^{(\text{est})} \times N(0,1) \tag{147}$$

   where $N(0,1)$ is a Gaussian (normal) variate with zero mean and unit variance.

3. You need to ensure that the MCMC is *efficient* — ideally, it will move from $x^t$ to the proposal value $x'$ about 1/3 of the time. Imagine if you had two highly degenerate parameters — say, $\Omega_M$ and $h$ in CMB measurements where only the combination $\Omega_M h^2$ is well determined. Say you use an otherwise reasonable proposal distribution which is a multivariate Gaussian with standard deviation equal to the *guessed* error in each parameter, and without correlation between parameters. Steps in $\Omega_M$ or $h$ separately will lead to rejection of the proposed steps vast majority of the time! However, if you are clever and reparameterize the problem so that you have parameter $\Omega_M h^2$ (and, say, $\Omega_M$ separately), then the acceptance will be much better, and the asymptotic distribution will be reached sooner. Equivalently, making the proposal function be an "off-diagonal" Gaussian with

   - directions specified by eigenvalues of the covariance matrix in the two parameters (which can be pre-computed with, for example, a short run to get the covariance)
   - parameter excursions equal to approximately 1-$\sigma$ steps along the eigenvectors – that is, steps are equal to (square roots of) the eigenvalues of the covariance matrix.

4. Finally, you need to make sure *mixing* of your chain. To do so, you can *thin* the chain, writing out every 100th (for example) value, so that you decrease the (otherwise very high) correlation between the steps. Likewise, you should run several (say, four) chains, and test convergence using one of the criteria (say, the *Gelman-Rubin criterion*) that typically compare variance within a chain with variance between different chains.

**MCMC: enjoying the fruits of labor.** MCMC is really a fantastic tool, enabling exploration of the multi-dimensional likelihoods that cannot be even contemplated using a naive multi-dimensional gridding of the parameter space.

Not only that, but computing constraints on any *derived* quantities of interest, once you have run your chains, is trivial as it can be done with *post-processing* of the MCMC output. What you need to do is write out chains, together with the "weight" (number of times the chain is "stuck" at that value if the proposed move was rejected) for each step. Then, for any parameter set of choice — a single parameter (e.g. $\Omega_M$), joint contour of two parameters (e.g. $(\Omega_M, w)$), a

function of a few parameters (e.g. $w(a) = w_0 + w_a(1-a)$, or the age of the universe $t_0$), whatever – you just look at their weights, rank order them, and add them until you get 68% or 95% or whatever fraction of the total weight. Remember, weight is proportional to posterior likelihood owing to ergodic property of MCMC.

Moreover, let us say that, after this hard work, that you decide you would like to combine your constraints with some other. That's easy — you just use constraints from your chain as a prior, and combine with new constraints to get a new posterior.

**Fisher information matrix: forecasting the errors.** Fisher matrix presents an excellent tool to *forecast* model parameter errors from a given experiment. Even though we have argued that the MCMC itself is "easy" and "fast" compared to brute force methods for exploring the likelihood, in comparison the Fisher matrix is still *much* easier and faster tool to forecast the likelihood distribution, given some expected experimental data.

Let us assume that we have cosmological measurements $\mathbf{X}$, and that the associated likelihood in the data can be represented by the likelihood $\mathcal{L}$. The Fisher matrix is formally defined as the curvature of the likelihood in model parameters $\mathbf{p}$ – that is, matrix of second derivatives of the log likelihood around the peak

$$\ln \mathcal{L}(\mathbf{p}) = \ln \mathcal{L}|_{\rm max} + \left.\frac{\partial \ln \mathcal{L}}{\partial p_i}\right|_{\rm max} (p_i - \bar{p}_i) + \frac{1}{2} \left.\frac{\partial^2 \ln \mathcal{L}}{\partial p_i \partial p_j}\right|_{\rm max} (p_i - \bar{p}_i)(p_j - \bar{p}_j) + \dots \quad (148)$$

$$= \ln \mathcal{L}|_{\rm max} + \frac{1}{2} \left.\frac{\partial^2 \ln \mathcal{L}}{\partial p_i \partial p_j}\right|_{\rm max} (p_i - \bar{p}_i)(p_j - \bar{p}_j) + \dots \quad (149)$$

where the summations are implied. The linear term vanishes, since the derivative at the maximum is zero! The Fisher matrix is now defined as the negative of the second derivative term (i.e. the *Hessian*) of the log likelihood:

$$\boxed{F_{ij} = \left\langle -\frac{\partial^2 \ln \mathcal{L}}{\partial p_i \partial p_j} \right\rangle} \quad (150)$$

where $\mathbf{p} \equiv \{p_i\}$ is the set of cosmological parameters. Basically, the Fisher matrix quantifies curvature around the peak of the likelihood. The higher the curvature, the better the parameters are determined, and the more information ("Fisher information") is available in the data regarding cosmological parameters.

When doing the Fisher matrix, we always assume that the likelihood in both the data $\mathbf{X}$ and in the cosmological parameters $\mathbf{p}$ is distributed as a multivariate Gaussian; then it follows that the covariance matrix of the data, $C$, has all information:

$$\mathcal{L} = \frac{1}{(2\pi)^{n/2} |\det C|^{1/2}} \exp\left[-\frac{1}{2}(X - \mu_i^T (C^{-1})_{ij} (X - \mu_j\right] \quad (151)$$

where $X_i$ are the data and $\mu_i$ are the theoretical observable quantities evaluated at the parameter values $\mathbf{p}$ at which we assume the Fisher contour is centered (e.g. $\mathbf{p} = \{\Omega_M = 0.3, \Omega_\Lambda = 0.7\}$). Also $C_{ij}$ is the covariance, and summation convention in $i$ and $j$ is employed. Then you could show as an exercise that the Fisher matrix evaluates to

$$\boxed{F_{ij} = \mu_{,i}^T C^{-1} \mu_{,j} + \frac{1}{2} {\rm Tr}[C^{-1} C_{,i} C^{-1} C_{,j}]} \quad (152)$$

where $_{,i}$ is partial derivative with respect to $p_i$.

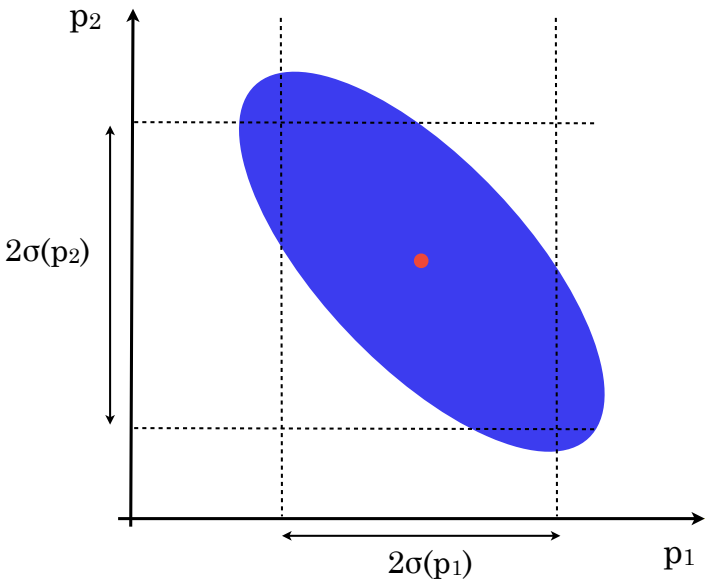


Figure 9: A sketch of how the likelihood contour in 2-dimensional plane is related to projected errors in the parameters. The figure shows a 68% confidence contour, obtained using the Fisher matrix (therefore it's a perfect ellipse by fiat).

In summary: if I assume the distribution in the cosmological parameters $\mathbf{p}$ is multi-variate Gaussian, then if I know how the curvature of the likelihood (i.e. how the data is related to model parameters), and if I also know errors in the data, I can forecast the expected errors in the model parameters. The Fisher matrix requires selecting the fiducial model (that is, central values of the parameters $\mathbf{p}$, as well as the assumption that the parameter-space likelihood is multivariate Gaussian.

**Fisher matrix as an estimate of parameter errors.** Most of the time, Fisher matrix users rely on the Cramer-Rao inequality (and theorem), which says that an error in a cosmological parameter $p_i$ will be greater or equal to the corresponding Fisher matrix element

$$\boxed{\sigma(p_i) \geq \begin{cases} \sqrt{(F^{-1})_{ii}} & \text{(marginalized)} \\ 1/\sqrt{F_{ii}} & \text{(unmarginalized)} \end{cases}} \tag{153}$$

where "marginalized" is the uncertainty marginalized over all other $N-1$ parameters, while the "unmarginalized" case is when you ignore the other parameters, assuming them effectively fixed and known. Note that the marginalized case has inverse of $F$ which lets the parameters "talk to each other" about degeneracies. Most often in cosmology we are interested in the marginalized errors; the unmarginalized ones are often much smaller and correspond to an unrealistic case when we somehow independently know the values of all other parameters.

So while the Cramer-Rao just tells us about the best possible error (so using the best possible estimator etc), we often just assume that it gives *the* error from data of the given quality.

**Examples.** Let us give some examples of the expressions for probe-specific Fisher matrices.

For type Ia supernova observations, the covariance matrix of SNe doesn't depend on cosmological parameters, and in fact it's often taken to be constant (remember, $C_{ij} \to \sigma_m^2 \delta_{ij}$ with

$\sigma_m \sim 0.15\,\mathrm{mag}$). Then Eq. (152) bevaluates to

$$F_{ij}^{\rm SNe} = \sum_{n=1}^{N_{\rm SNe}} \frac{1}{\sigma_m^2} \frac{\partial m(z_n)}{\partial p_i} \frac{\partial m(z_n)}{\partial p_j} \tag{154}$$

where $m(z) = m(z, \Omega_M, \Omega_\Lambda, \mathcal{M}...)$ is the theoretically expected apparent magnitude. Notice that, if you had a full off-diagonal covariance matrix $C_{ij}$ that is still independent of the cosmological parameters, the equation above would generalize trivially.

For the counts of galaxy clusters (counting them in mass and redshift, and comparing to theory), the Fisher matrix can be calculated as follows. Let $N_k$ be number of clusters in the $k$th bin and $O(z)$ be an observable (say, X-ray flux), then

$$F_{ij}^{\rm clus} = \sum_{k=1}^{Q} \frac{N_k}{\sigma_O(z_k)^2} \frac{\partial O(z_k)}{\partial p_i} \frac{\partial O(z_k)}{\partial p_j} \tag{155}$$

Now consider a case of measurements of the CMB (or weak lensing) power spectrum where the mean (temperature or shear) is zero and doesn't depend on cosmology, but the covariance carries all cosmological information. Then you can show that

$$F_{ij}^{\rm WL} = \sum_{\ell} \frac{\partial C^\kappa(\ell)}{\partial p_i} \mathbf{Cov}^{-1} \frac{\partial C^\kappa(\ell)}{\partial p_j}, \tag{156}$$

where $\mathbf{Cov}^{-1}$ is the inverse of the covariance matrix between the observed power spectra whose elements are given by

$$\mathrm{Cov}\left[C_{ab}^\kappa(\ell), C_{cd}^\kappa(\ell)\right] = \frac{\delta_{\ell\ell'}}{(2\ell+1)\, f_{\rm sky}\, \Delta\ell} \left[C_{ac}^\kappa(\ell) C_{bd}^\kappa(\ell) + C_{ad}^\kappa(\ell) C_{bc}^\kappa(\ell)\right]. \tag{157}$$

where $C_{ab}^\kappa(\ell)$ is the covariance of convergence $\kappa$ between galaxies in the $a$-th and $b$-th redshift bin, on scales corresponding to a multipole bin centered at $\ell$ with width $\Delta\ell$, in a survey covering $f_{\rm sky}$ fraction of the sky. You can tell by eyeballing this covariance-of-covariance four-point correlation function that it was computed using Wick's theorem, that is, assuming Gaussianity of $C$.

**Marginalization over parameters.** If you have, say, $N$, cosmological parameters, how do you marginalize over $N-M$ of them to be left with a desired joint constraints on $M$ parameters? This is easy:

- Calculate the full $N \times N$ matrix $F$
- Invert it to get $F^{-1}$
- Take the desired $M \times M$ subset of $F^{-1}$, and call it $G^{-1}$ note that this matrix is $M$ dimensional
- Invert $G^{-1}$ to get $G$

and voilà — the matrix $G$ is the projected Fisher matrix onto the $M$-dimensional space. Notice that the step of inverting $F$ assures that the parameters "talk to each other", which effectively accounts for the marginalization (and increases the error bars significantly compared to the unmarginalized error).

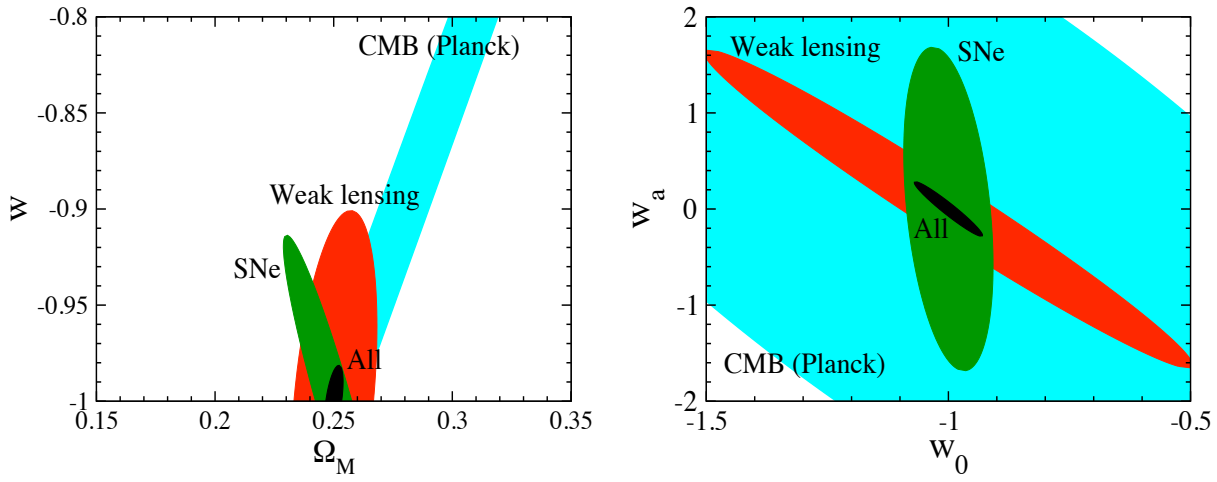


Figure 10: Illustration of forecast constraints on dark energy parameters, taken from the Frieman, Turner & Huterer review of DE. All contours have been computed using the Fisher matrix. Note that the contours are by definition ellipses, and one (Planck) is nearly completely degenerate — meaning, long.

**Fisher ellipses.** How do you plot the Fisher matrix contour? To plot a 2D ellipse, you first want to project down to that space, and be left with a marginalized 2x2 Fisher matrix, call it G. The equation for the 2D ellipse is

$$G_{11}p_1^2 + 2G_{12}p_1p_2 + G_{22}p_2^2 = \Delta\chi^2_{2\,\mathrm{dof}} \equiv \frac{1}{f} \tag{158}$$

where. for two parameters (i.e. for 2-D ellipse), $f = 0.434$ for a 68% CL ellipse, and $f = 0.167$ for a 95% ellipse (these numbers can easily be calculated using the $\chi^2$ statistic, and you will produce them on your homework!). More generally, the equation of the n-dimensional ellipsoid (that is, with $n$ dof) would be

$$\boxed{(\mathbf{p}-\bar{\mathbf{p}})^T F\,(\mathbf{p}-\bar{\mathbf{p}}) = \Delta\chi^2_{n-\,\mathrm{dof}}} \quad \text{(equation of Fisher ellipsoid)}. \tag{159}$$

The Fisher ellipse are also show how much information is carried by the data on the parameters — the smaller the ellipse, the more information. More generally, the volume of an $n$-dimensional ellipsoid is

$$\boxed{\text{Volume} \propto (\det F)^{-1/2}} \quad \text{(volume of Fisher ellipsoid)}. \tag{160}$$

You can find this useful if you are estimating relative amounts of information in surveys, etc.

**Fisher bias.** Another great application of the Fisher matrix is to calculate the bias in parameters $p_i$ given biases in the observables. This can be derived easily again assuming the same Gaussian distribution in the data; the result (for weak lensing) is

$$\delta p_i \approx F_{ij}^{-1} \sum_\ell \left[C^\kappa_\alpha(\ell) - \bar{C}^\kappa_\alpha(\ell)\right]\, \mathrm{Cov}^{-1}\left[\bar{C}^\kappa_\alpha(\ell), \bar{C}^\kappa_\beta(\ell)\right]\, \frac{\partial \bar{C}^\kappa_\beta(\ell)}{\partial p_j}, \tag{161}$$

where $[\bar{C}^\kappa_\alpha(\ell), \bar{C}^\kappa_\beta(\ell)]$ is the bias in the "observable" shear covariance due to any reason.

Perhaps a simpler example would be that for SNeIa, where the bias in the parameters takes the form

$$\delta p_i \approx F_{ij}^{-1} \sum_n \frac{1}{\sigma_m^2} \left[ m(z_n) - \bar{m}(z_n) \right] \frac{\partial \bar{m}(z_n)}{\partial p_j} \tag{162}$$

where $[m(z_n) - \bar{m}(z_n)]$ is the bias in the observed apparent magnitudes.

The bias formula is extremely useful if you would like to see what effect on cosmological parameter errors an arbitrary systematic effect makes. So, given some biases in the observable quantities, for example $[C_\alpha^\kappa(\ell) - \bar{C}_\alpha^\kappa(\ell)]$ in Eq. (161), you can find biases in the cosmological parameters $\delta p_i$. Then you can compare those biases with the statistical errors in the cosmological parameters $\sigma(p_i)$ and impose requirements on the control of your systematic effect so that $|\delta p_i|/\sigma(p_i)$ is no larger than some threshold, say 0.3 (corresponding to $< 30\%$ bias in the parameters). See the paper by Huterer & Takada (2006) for application to how well theoretical prediction for the power spectrum $P(k)$ needs to be know in order not to "mess up" cosmol