Optimal combination of the spatial correlation function and counts of galaxy clusters

2-Months Summer Internship Report, CEA Paris-Saclay

Yoann Launay[a,b] **and Jean-Baptiste Melin**[c]

[a]University of Cambridge; [b]CentraleSupélec at Paris-Saclay University; [c]Irfu, CEA Paris-Saclay

(Rina Piccolo)

1

Table of Contents

Introduction

2

To understand the story of the universe, physicists have identified cosmological observables : the CMB, supernovae but also galaxy clusters hosting most of the matter in the universe. From these last ones, many direct and indirect observables can be reconstructed including galaxy clusters counts and space correlations. Between these two, correlations are expected and encourage the exploration of a unification of two distinct formalisms - at least in the refinement of cosmological parameters. This is the goal of this internship, the results of which are exposed in the following after some personal and brief elements of comprehension of the subject.

In the following, we expect the reader to have necessary skills and cosmology's essentials in mind.

A github repository is also accessible at the following link, with any program which was used in this report.

1. Counting Clusters

18

A. Essential theory of independent events. *The major part of the following theoretical elements is a summary of (1).*

Galaxy clusters represent great amounts of matter in the universe, especially dark matter. In this way, this is where gravitational fields are non-negligeable. In fact, after the decoupling of the universe at redshift 1100, the only rival phenomenon of the gravity is the expansion of the universe. As a first instinct, we all have this image of a balloon with linked balls on it. Links, balls and the growing balloon respectively being gravity attraction, massive objects and expanding universe, the question is : when do we have the right to create a link ? Physicists have thus expected a threshold between a scale of linked distributions of matter regulated by gravity and a universe scale where gravitational effects are negligeable in the metric behind the expansion rate. This is the reason

Correspondence: Yoann Launay yoann.launay@outlook.com

why the story of matter is often characterized as an accretion of the primordial matter distribution or what could be called a "cosmic web".

The Halo model is today the main theory allowing to predict counts of galaxy clusters in a given area of the sky. **The associated scale is the one of a galaxy cluster**. The main results are briefly reminded in the following.

- It starts with the energy balance of a supposed spherical cluster system of mass M and (physical) radius r from t_i singularity time to t time:

$$\frac{1}{2}(\frac{dr}{dt})^2 - \frac{GM}{r} - \frac{H^2\Omega_v r^2}{2} = \frac{1}{2}(H_i r_i)^2 - \frac{GM}{r_i} - \frac{H^2\Omega_v r_i^2}{2} \quad [1]$$

The crucial point is then to determine what is called the "turnaround radius" r_{ta}, known solution of $(\frac{dr}{dt})^2 = 0$: this is the time when the expansion stops its prominence over gravity and from which it can decrease until the collapse of the Halo (inverse evolution and so $t_{collapse} = 2t_{ta}$. However the massive object tends to an hydrostatic equilibrium before the collapse : virialization occurs in fact when Viriel Theorem holds ($2 * E_c + V_G = 0$). Figure 1 illustrates the evolution of the physical radius of the massive object.

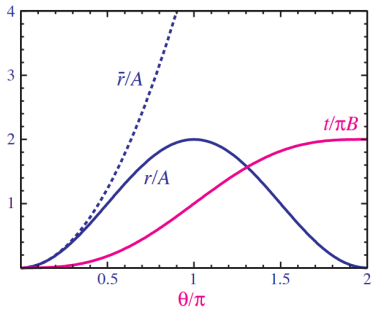

Fig. 1. Parametric (in θ) solutions for the radius ($r(\theta) = A(1 - cos(\theta))$ and $t(\theta) = B(\theta - sin(\theta))$). Turnaround and collapse respectively at "time" $\theta = \pi$ and 2π. (Wayne Hu)

- The over-density contrast or fluctuation density field is the common tool used to describe massive terms in the previous resolution. Its definition depends on the mean density of the universe $\overline{\rho}$ and the position $\vec{x}$ of the cluster:

$$\delta(\vec{x}, t) = \frac{\rho_(\vec{x}, t) - \overline{\rho}}{\overline{\rho}} \quad [2]$$

The evolution equation of this field is usually derived and linearized from Fluid Mechanics's Continuity and Euler equations and Friedmann density equations but also from gravitational Poisson equation. Then a separation of variables follows the linear approximation and we get $\delta(\vec{x}, t) = D_+(t)\delta(\vec{x})$ with $\delta(\vec{x})$ the primordial fluctuation density field and D_+ the linear growth factor following a second order equation with a well-established integrative solution depending on the cosmological parameters.

Usually, the local version of δ in the halo model scale is simply denoted $\delta(t)$.

- There exists then a critical fluctuation density field δ_c which is reached at the collapse of a forming cluster. It is quite common not to take the explicit formula to compute it because it does not vary a lot with parameters and redshift (time) : a 1.686 value coming from Einstein-De Sitter universe is a good reference as a constant.

- Then the number of expected virialized objects per volume units can be written with probability theory. If we consider that at each spatial point, the same independent probability to be virialized at redshift z (or time t) and mass M occurs and that it corresponds to the fraction of fluctuations not less than δ_c at redshift z, one can write the mass function (2):

$$\mathbb{P}(\delta > \delta(z_i)) = \int_{\delta(z_i)=\delta_c(z)\frac{D_+(z_i)}{D_+(z)}} \frac{1}{\sqrt{2\pi}\sigma(M, z_i)} e^{\frac{-\delta^2}{2\sigma(M,z_i)^2}} d\delta \quad [3]$$

We have in fact supposed that δ is gaussian random process in space with standard deviation σ called root-mean-square (RMS) density fluctuation. This is the standard deviation of a gaussian process with zero mean but considering densities smoothed at a given scale (a radius or the equivalent mass for a given equivalence relationship) :

$$\sigma(R, t) = \sqrt{< \delta_R(\vec{x}, t)^2 >} = D_+(t)\sqrt{< \delta_R(\vec{x})^2 >} \quad [4]$$

where the smooth field in a V_R volume writes $\delta_R = \frac{1}{V_R}\int_{V_R} \delta(\vec{x} + \vec{r})d^3\vec{r}$.

- Nevertheless, the Fourier space allows us to explicit the formula with the primordial power spectrum $P(k)$ (and at any time with a D_+^2 factor) of the density field given by inflation theory (3) and thanks to the product (respectively convolution in real space) with the R-scale smoothing function W (respectively with w = $\frac{\mathbb{I}_{||\vec{x}||\leq R}}{4\pi R^3/3}$) :

$$\sigma(R)^2 = \frac{1}{(2\pi)^3}\int P(k)|W(kR)|^2 d^3\vec{r} \quad [5]$$

The normalization of such a spectrum is made thanks to the value of what is now considered as a cosmological parameter : σ_8 as the primordial RMS density fluctuation at $8h^{-1}$Mpc.

- Then the distribution of matter in time, mass and isotropic universe is deduced from the probability of equation 3. The number of virialized objects per volume unit and with mass between M and dM is supposed to be the density number of virialized regions weighted by the probability of a virialized object to be at the given mass interval:

$$dn = \frac{\overline{\rho}}{M} \times |\frac{d\mathbb{P}}{dM}|dM \quad [6]$$

This how we find back Press-Schechter Halo Mass Function giving the distribution of clusters in mass and redshift space (with factor 2 normalization correction) :

$$\frac{dn}{d\ln M}(M, z) = \sqrt{\frac{2}{\pi}}\frac{\overline{\rho}}{M}\left|\frac{d\ln\sigma}{d\ln M}(M, z)\right| \nu(M, z)e^{-\frac{1}{2}\nu(M,z)^2} \quad [7]$$

with $\nu(M, z) = \frac{\delta_c(z)}{\sigma(M,z)}$.

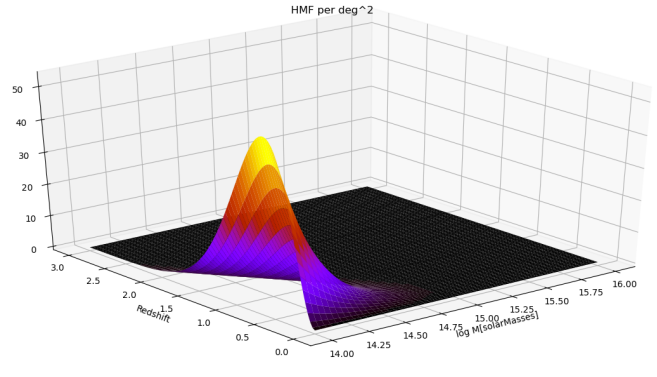

Fig. 2. Halo Mass function per squared degree and redshift ($\frac{dN}{d\ln M dz}(M,z)$).

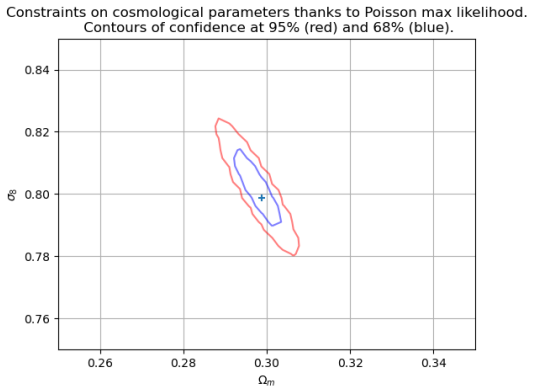

Fig. 3. Reproduction of Planck 2015 Collaboration work : refinement of the Poisson log-likelihood on randomly Poisson-generated data ($z_{min} = 0$, $z_{max} = 3$, $M_{min} = 5\times10^{14}M_\odot$, $M_{max} = 1\times10^{16}M_\odot$, $\Omega = 4\pi$). A Gaussian likelihood give the same result.

We now see how it can predict counts in a portion of sky since the distribution of clusters is now given by the integration of the number of clusters per *steradian*, redshift and mass interval depending on the previous distribution and the differential comoving volume :

$$\frac{dN}{d\ln M d\Omega dz} = \frac{dn}{d\ln M}\frac{dV}{d\Omega dz} \quad [8]$$

This distribution is represented on figure 2 : one can identify both the typical redshift peak (around 1) and orders of magnitudes ($1e14$ to $1e15$ solar masses) of the mass of virialized clusters.

B. Refinement and cosmological constraints. If we denote Θ a cosmology as a set of cosmological parameters (Ω_m,Ω_v, σ_8 etc) we have a clear dependance in the previous expressions :

$$\frac{dN}{d\ln M d\Omega dz}(\Theta) = \frac{dn}{d\ln M}(\Theta)\frac{dV}{d\Omega dz}(\Theta) \quad [9]$$

For a given bin we can then predict the expected number of counts : $\overline{N_{kl}}(\Theta) = \Omega\int_{\ln M_k}^{\ln M_{k+1}}\int_{z_l}^{z_{l+1}}\frac{dN}{d\ln M d\Omega dz}(\Theta)$

Most approaches (4) consider then this modeled value as the mean value of a univariate Poisson distribution in space. This is why for given bins, the log-likelikood which is refined in practice on data counts (N_{kl}) is :

$$\ln\mathcal{L}_N(\Theta) = \sum_{k,l} -\overline{N_{kl}}(\Theta) + N_{kl}\ln(\overline{N_{kl}}(\Theta)) - \cancel{\ln(N_{kl}!)} \quad [10]$$

The Planck 2015 collaboration has then allowed to get new constraints on the cosmology (5) with this method, like in figure 3.

However these counts suppose independent bins as the likelihood probability and more generally the halo model considers probabilities to be the same for any point and without correlations.

In many other domains, physicists (6) have given credit to a multivariate Poisson distribution (7, 8). However this distribution can really become hard to compute and demands great performances but also rigor since many versions of this interdependence exist.

Hopefully, since sufficient scales allow to consider many counts, it is quite useful and common to approach the Poisson laws to Gaussian ones. This is in this case easier to express the correlations directly with a covariance matrix if bins are now considered dependent in the Gaussian log-likelihood :

$$\ln\mathcal{L}_N(\Theta) = -\frac{1}{2}(\overline{N}(\Theta)-N)^T\Sigma_N^{-1}(\overline{N}(\Theta)-N) \quad [11]$$

with $\overline{N}$ and N column matrices respectively containing the $(\overline{N_{kl}})_{k,l}$ and the $(N_{kl})_{k,l}$, and with $diag(\Sigma_N)$ usually equal to the Poisson variances of $\overline{N}$ and extra-diagonal terms as convolution of smoothing functions and of the power spectrum (9). It is often stated as negligible as it was showed before (9) except for specific surveys.

2. Space correlation function

The major part of the following theoretical elements is a summary of (1) and (10).

In Euclid mission, new scales will be explored in catalogs and will demand a full consideration of correlations between bins. Those catalogs will be workable with 2 formalisms: cluster counts and space correlations as were suggesting last discussions of the previous section.

A. Essential theory of dependent events. The space correlation function also uses the formalism of the initial fluctuation density field (going to time t with D_+) but fully considering it as a 3D-spatial gaussian process in its stochastic meaning:

- If δ is a random variable at each point, the correlation is also as a mean of these values over a finite volume V:

$$\xi(\vec{r}) =< \delta(\vec{x})\delta(\vec{x}+\vec{r}) >= \frac{1}{V}\int_V \delta(\vec{x})\delta(\vec{x}+\vec{r})d^3\vec{x} \quad [12]$$

  Mathematically, it is usual for the correlation to be confused with the covariance of a random variable field. However, this definition would not allow ξ to be a random variable too.

- Peebles and Mellier formalism (10) takes a physical origin for the definition of the correlation. This is the excessive crossed probability density to find two massive volumes compared to a uniform case :

$$dN(\vec{x_1},\vec{x_2}) = \overline{\rho}^2(1+\xi(\vec{r_{12}}))dV_1dV_2 =< \rho(\vec{x_1})\rho(\vec{x_2}) > dV_1dV_2 \quad [13]$$

- This link with δ brings to mind that their power spectrum is the same, useful for any analytical computation :

$$\xi(\vec{r}) \propto \int P(k) e^{i\vec{k}\cdot\vec{r}} d^3\vec{k} \quad [14]$$

where we can explicit this famous initial power spectrum model coming from primordial inflation theory:

$$P(k) = A_s k^{n_s} T(k)^2 \quad [15]$$

with A_s the normalization computed from σ_8 or reversely, n_s the spectral index of fluctuation, T the transfer function depending on Ω_m and Ω_v.

These equations allow to write a completely Θ-dependent analytical expression of the space correlation function (often considered isotropic in this context) : $\xi(r, \Theta)$.

B. Refinement and cosmological constraints. How to construct data to refine this model on ? The construction of such an indirect observable starts when assuming the mass of each cluster being its constant mean value. The method is then to count pairs of galaxy clusters and to evaluate their distance to each others. The use of a random Catalog is however necessary to keep a simple expression and to stabilize the result. The common tool is the Landy and Szaslay estimator (11) :

$$\widehat{\xi}(r) = \frac{[DD]_r - 2[DR]_r + [RR]_r}{[RR]_r} \quad [16]$$

where D and R denote respectively the studied catalog of clusters with their coordinates and a uniform random one, $[AB]_r$ notations representing the number of pairs with interdistance of $r \pm dr/2$ (dr defining comoving distance bin widths).

Then a univariate gaussian log-likelihood as in the following expression allows to get new constraints on the cosmological parameters (see figure 8).

$$\ln \mathcal{L}_\xi(\Theta) = -\frac{1}{2}(\xi_i(\Theta) - \widehat{\xi})^T \Sigma_\xi^{-1} (\xi_i(\Theta) - \widehat{\xi}) \quad [17]$$

with ξ and $\widehat{\xi}$ the column matrices for $(\xi_i)_i$ (as theoretical points, see eq. 14) and $(\widehat{\xi_i})_i$ (Landy and Szaslay estimator, see eq. 16) , and with $\Sigma_{\xi,ii} = \sigma^2(\widehat{\xi_i})$. In this expression, it is in fact assumed too that non-diagonal terms of the covariance matrix are negligible (see example figure 4).

3. Toward correlating counts

From now, this paper introduces personal research on the subject.

A. Introductive theoretical aspects. Following the teachings of the construction of Landy and Szaslay or Peebles estimators, one can immediately formalize the link between counts and the correlation function thanks to the first approximation of a cluster mass equal to its mean value for each galaxy cluster :

$$1 + \widehat{\xi}(r) = \frac{\overline{M}^2_{cluster}}{\overline{\rho}^2 V^2_{bin}} < N(x)N(x+r) > \quad [18]$$

As a simpler version, let be (N_l) random variables representing counts of bins in a 1D space. We can then write the 1D correlation as a random variable for a separation of n dr-separated bins:

$$1 + \widehat{\xi}(n \times dr) \propto lim_{w\to\infty} \frac{1}{2w+1} \sum_{k=-w}^{w} N_k N_{k+n} \quad [19]$$

Because this limit is supposed finite, the covariance of it with counts of any bin m writes :

$$Cov(\widehat{\xi}(n\times dr), N_m) \propto lim_{w\to\infty} \frac{1}{2w+1} \sum_{k=-w}^{w} Cov(N_k N_{k+n}, N_m) \quad [20]$$

Let suppose now that counts are independent random variables. Then the number of non-zero terms in this sum is finite (at most 2 terms): the limit and so $Cov(\widehat{\xi}(n \times dr), N_m)$ is zero for any m. However, considering the (N_m) as a basis set in the abstract Hilbert space of random variables with Cov scalar product, we deduce that such an equality implies the nullity of $\widehat{\xi}(n \times dr)$. That means an absence of correlation if counts are independent and I think that this 1D case generalises.

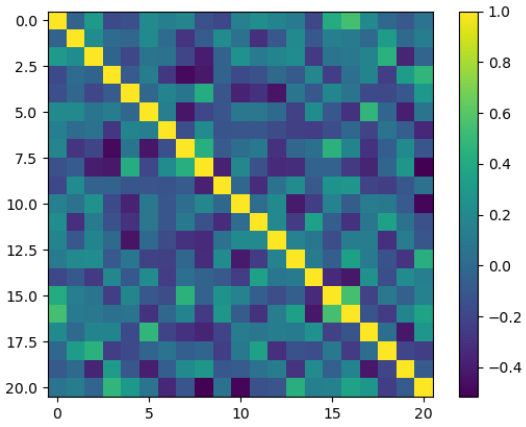

Fig. 4. An example of Σ_ξ matrix as the covariance matrix of the Landy and Szaslay estimators of 10 catalogs generated by the method explained in section C.

B. Unified Likelihood. Counts and space correlation formalisms have many common points but also seem to be correlated. With a given catalog of objects, one could then maximize the product (since $\mathcal{L}$ is by definition a probability) of the previous gaussian likelihood, assuming correlated counts (see section 3.A for justification, assuming then a full Σ_N) and more generally correlated correlation bins (through a full Σ) :

$$\ln \mathcal{L} = \ln \mathcal{L}_N + \ln \mathcal{L}_\xi$$

This is strictly equivalent to a unique Gaussian multivariate log-likelihood :

$$\ln \mathcal{L} = -\frac{1}{2}(X(\Theta) - X^{exp})^T \Sigma^{-1} (X(\Theta) - X^{exp}) \quad [21]$$

with the concatenated column matrices $X = ((\overline{N}_{kl}), (\xi_i))$ and $X^{exp} = ((N_{kl}), (\widehat{\xi_i}))$ and the global covariance matrix :

$$\Sigma = \begin{pmatrix} \Sigma_N & \Sigma_{N\xi}^T \\ \Sigma_{N\xi} & \Sigma_\xi \end{pmatrix} \quad [22]$$

where $\Sigma_{N\xi}$ is assumed to be equal to zero. Nevertheless, this assumption has to be tested and each part of Σ must be theoretically expected. This is the goal of this section, considering both ξ and δ as random variable fields (each point is a random variable and thus our universe is only one realization of this random field). In the following, any bin is denoted by its centered position ($\vec{x}$ for instance).

B.1. Counts-counts covariance. The non-diagonal terms of Σ_N have been modelized (9) starting from the definition of the counts density :

$$n_{\vec{x}} = \int \overline{n}(1+\delta(\vec{x_i}))w_{R,\vec{x}}(\vec{x_i})d^3\vec{x_i} = \overline{n}(1+\delta_R(\vec{x})) \quad [23]$$

In fact, it comes then that :

$$\begin{aligned} Cov(\tfrac{n_{\vec{x}}}{\overline{n}}, \tfrac{n_{\vec{y}}}{\overline{n}}) = &\ Cov[\delta_R(\vec{x}), \delta_R(\vec{y})] \\ = &\ \mathbb{E}[\delta_R(\vec{x}), \delta_R(\vec{y})] \\ = &\ \mathbb{E}[\int\int \delta(\vec{x_i})w_{R,\vec{x}}(\vec{x_i})\delta(\vec{x_j})w_{R,\vec{y}}(\vec{x_j})d^3\vec{x_i}d^3\vec{x_j}] \\ = &\ \int\int w_{R,\vec{x}}(\vec{x_i})w_{R,\vec{y}}(\vec{x_j})\mathbb{E}[\delta(\vec{x_i})\delta(\vec{x_j})]d^3\vec{x_i}d^3\vec{x_j} \end{aligned} \quad [24]$$

However, a strong assumption holds in (9) since sample variables are considered, identifying the expectation and the observed mean. Over a great V volume in front of the bins, one could in fact write :

$$\xi(\vec{x_j} - \vec{x_i}) \simeq \mathbb{E}[\delta(\vec{x_i})\delta(\vec{x_j})] \quad [25]$$

And thus :

$$\xi_R(\vec{y} - \vec{x}) \simeq Cov(\frac{n_{\vec{x}}}{\overline{n}}, \frac{n_{\vec{y}}}{\overline{n}}) \quad [26]$$

This ergodic assumption is typically withdrawing the status of the $(\xi(\vec{r}))_{\vec{r}}$ as non-constant random variables and this is not relevant if one wants to compute theoretical values of $\Sigma_{N\xi}$ or Σ_ξ. This is why section B.2 starts from a random variable for ξ_R (see eq. 27).

B.2. Correlation-counts covariance. If eq. 18 gives a clear dependence between ξ and counts, it stays integral and we must work with the R-convoluted version of both before computing the covariance :

$$\begin{aligned} \xi_R(\vec{r}) = &\ < \delta_R(\vec{x})\delta_R(\vec{x}+\vec{r}) > \\ = &\ \tfrac{1}{V}\int_V d^3\vec{x}\int\int \delta(\vec{x_i})w_{R,\vec{x}}(\vec{x_i})\delta(\vec{x_j})w_{R,\vec{x}+\vec{r}}(\vec{x_j})d^3\vec{x_i}d^3\vec{x_j} \end{aligned} \quad [27]$$

Then, the binned counts densities appear when introducing 23 :

$$\xi_R(\vec{r}) = \frac{1}{V}\int_V \left(\frac{n_{\vec{x}}}{\overline{n}} - 1\right)\left(\frac{n_{\vec{x}+\vec{r}}}{\overline{n}} - 1\right) d^3\vec{x} \quad [28]$$

We now have a convenient expression to compute the covariance of interest between a correlation point $\xi_R(\vec{r})$ and a mean-normalized counts density $\frac{n_{\vec{y}}}{\overline{n}}$ in bin $\vec{y}$:

$$Cov(\xi_R(\vec{r}), \tfrac{n_{\vec{y}}}{\overline{n}}) = \ \tfrac{1}{V}\int_V Cov(\left(\tfrac{n_{\vec{x}}}{\overline{n}} - 1\right)\left(\tfrac{n_{\vec{x}+\vec{r}}}{\overline{n}} - 1\right), \tfrac{n_{\vec{y}}}{\overline{n}})d^3\vec{x} \quad [29]$$

The first term can be explicited thanks to the following development:

$$\begin{aligned} Cov(XY, Z) = &\ \mathbb{E}[X]Cov(Y,Z) + \mathbb{E}[Y]Cov(X,Z) \\ &+\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])(Z - \mathbb{E}[Z])] \end{aligned} \quad [30]$$

So that, replacing X, Y and Z random variables by the binned densities contrast with zero mean, we have :

$$Cov(\frac{n_{\vec{x}}n_{\vec{x}+\vec{r}}}{\overline{n}^2}, \frac{n_{\vec{y}}}{\overline{n}}) = \mathbb{E}[(\frac{n_{\vec{x}}}{\overline{n}} - 1)(\frac{n_{\vec{x}+\vec{r}}}{\overline{n}} - 1)(\frac{n_{\vec{y}}}{\overline{n}} - 1)] \quad [31]$$

Hence re-writing 29 as :

$$\begin{aligned} Cov(\xi_R(\vec{r}), \tfrac{n_{\vec{y}}}{\overline{n}}) = &\ \tfrac{1}{V}\int_V \mathbb{E}[(\tfrac{n_{\vec{x}}}{\overline{n}} - 1)(\tfrac{n_{\vec{x}+\vec{r}}}{\overline{n}} - 1)(\tfrac{n_{\vec{y}}}{\overline{n}} - 1)]d^3\vec{x} \\ = &\ \tfrac{1}{V}\int_V \mathbb{E}[\delta_R(\vec{x})\delta_R(\vec{x}+\vec{r})\delta_R(\vec{y})]d^3\vec{x} \\ = &\ \mathbb{E}\left[\tfrac{1}{V}\int_V \delta_R(\vec{x})\delta_R(\vec{x}+\vec{r})\delta_R(\vec{y})d^3\vec{x}\right] \end{aligned} \quad [32]$$

This result could have been found quicker by directly applying König-Huygens formula to compute the covariance.

In any case, the 3-points correlation function (3PCF) seems to appear in the expectation (under some symmetries assumptions and considering a fixed point in space), while ξ is the 2PCF.

B.3. Correlation-correlation covariance. While it is possible to find this covariance through the one of the power spectrum (12), we propose here a try starting from 28 :

$$\begin{aligned} Cov[\xi_R(\vec{r_1}), \xi_R(\vec{2})] = &\ \tfrac{1}{V^2}\int_V\int_V Cov[\left(\tfrac{n_{\vec{x}}}{\overline{n}} - 1\right)\left(\tfrac{n_{\vec{x}+\vec{r_1}}}{\overline{n}} - 1\right) \\ &, \left(\tfrac{n_{\vec{y}}}{\overline{n}} - 1\right)\left(\tfrac{n_{\vec{y}+\vec{r_2}}}{\overline{n}} - 1\right)]d^3\vec{x}d^3\vec{y} \end{aligned} \quad [33]$$

With the same approach as in B.2, and knowing that $\mathbb{E}[\frac{n_{\vec{x}}}{\overline{n}} - 1] = 0$, we use :

$$\begin{aligned} Cov[AB, CD] = &\ \mathbb{E}[ABCD] - \mathbb{E}[AB]\mathbb{E}[CD] \\ = &\ \mathbb{E}[ABCD] \\ &-(Cov[A,B] + \mathbb{E}[A]\mathbb{E}[B])(Cov[C,D] + \mathbb{E}[C]\mathbb{E}[D]) \end{aligned} \quad [34]$$

so that :

$$\begin{aligned} Cov[\xi_R(\vec{r_1}), \xi_R(\vec{r_2})] = &\ \tfrac{1}{V^2}\int_V\int_V \mathbb{E}[\left(\tfrac{n_{\vec{x}}}{\overline{n}} - 1\right)\left(\tfrac{n_{\vec{x}+\vec{r_1}}}{\overline{n}} - 1\right) \\ &\times \left(\tfrac{n_{\vec{y}}}{\overline{n}} - 1\right)\left(\tfrac{n_{\vec{y}+\vec{r_2}}}{\overline{n}} - 1\right)] \\ &-Cov[\tfrac{n_{\vec{x}}}{\overline{n}} - 1, \tfrac{n_{\vec{x}+\vec{r_1}}}{\overline{n}} - 1] \\ &\times Cov[\tfrac{n_{\vec{y}}}{\overline{n}} - 1, \tfrac{n_{\vec{y}+\vec{r_2}}}{\overline{n}} - 1]d^3\vec{x}d^3\vec{y} \\ = &\ \mathbb{E}\left[\tfrac{1}{V^2}\int_V\int_V \delta_R(\vec{x})\delta_R(\vec{x}+\vec{r_1}) \right. \\ &\left. \times\delta_R(\vec{y})\delta_R(\vec{y}+\vec{r_2})d^3\vec{x}d^3\vec{y}\right] \\ &-\tfrac{1}{V^2}\int_V\int_V Cov[\tfrac{n_{\vec{x}}}{\overline{n}} - 1, \tfrac{n_{\vec{x}+\vec{r_1}}}{\overline{n}} - 1] \\ &\times Cov[\tfrac{n_{\vec{y}}}{\overline{n}} - 1, \tfrac{n_{\vec{y}+\vec{r_2}}}{\overline{n}} - 1]d^3\vec{x}d^3\vec{y} \\ = &\ \mathbb{E}\left[\tfrac{1}{V^2}\int_V\int_V \delta_R(\vec{x})\delta_R(\vec{x}+\vec{r_1}) \right. \\ &\left. \times\delta_R(\vec{y})\delta_R(\vec{y}+\vec{r_2})d^3\vec{x}d^3\vec{y}\right] \\ &-\tfrac{1}{V^2}\int_V\int_V Cov[\tfrac{n_{\vec{x}}}{\overline{n}} - 1, \tfrac{n_{\vec{x}+\vec{r_1}}}{\overline{n}} - 1] \\ &\times Cov[\tfrac{n_{\vec{y}}}{\overline{n}} - 1, \tfrac{n_{\vec{y}+\vec{r_2}}}{\overline{n}} - 1]d^3\vec{x}d^3\vec{y} \end{aligned} \quad [35]$$

This is also just a re-writing thanks to König-Huygens formula, allowing to see the 4PCF (under symmetries assumptions and points taken as constants).

C. Unifying catalogs. In practice, we cannot use two different catalogs for correlations and counts and constitute a simple Gaussian likelihood to refine. The source of data has to be the same for both. Moreover necessary scales to maximize the induction of co-effects between counts and correlations are those of Euclid mission, which are not available yet. This is why simulations will help us.

C.1. Density field Generator. In this context, it was decided to simulate a box of $n_c^3 = 256^3$ cubic pixels representing $dx =$ 20Mpc primordial comoving cubes. This is approximately both the $8h^{-1}$Mpc smoothing scale and the cluster one.

It is in fact possible to simulate a δ field in this box by imposing the $P(k)$ spectrum to an initial random Gaussian noise, by passing through the Fourier space (13).

In fact, if (δ^G_{ijk}) is a 3D-finite random univariate Gaussian field, the FFT allows to get an approximation of the discrete Fourier transform :

$$\Delta_{lmn} = \mathcal{F}\{(\delta^G_{ijk})\} \quad [36]$$

Then, imposing the known theoretical spectrum $P(k)$ demands :

$$\Delta^{(1)}_{lmn} = \Delta_{lmn}\sqrt{P_{lmn}} \quad [37]$$

A random primodrial universe is then obtained thanks to the inverse 3D-Fourier transform :

$$\delta_{ijk} = \mathcal{F}^{-1}\{(\Delta^{(1)}_{lmn})\} \quad [38]$$

At this stage, it is already possible to check the correlation function in this simulated box. The best way we found to evaluate it without creating any geometric or convoluting bias was drawing continuous positions and computing the product of densities for any two of them. While the positions were continuous, their densities were those of the cubic cell they were attached to. A demonstrative box is plotted on figure 5 and was obtained thanks to the *gen.py* code.

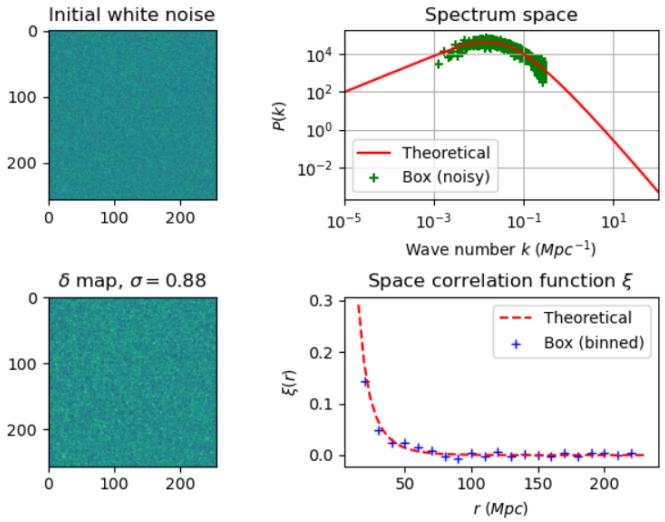

Fig. 5. Results of a 256^3 simulated box

We notice here the effect of the pixellisation since the effective σ is 0.88 and not 0.80 as imposed in the spectrum normalization. This is due to the choice of dx which is not exactly $8h^{-1}$Mpc on the one hand, but also on the other hand pixels are not spherical. Nevertheless taking the appropriate dx can give $\sigma = \sigma_8$ but in any case the box has the correct value for σ_8 as imposed by the spectrum construction.

C.2. Draw of catalogs.

Draw from probability density function A real catalog is not a field of δ values. It is a list of objects detected at different positions in the celestial sphere and potentially with its mass.

A naive method to draw such a catalog is to take a threshold for δ and take a cluster at each pixel over the threshold. However this is not a stochastic approach and it destructs the spectrum, as many of our tests showed.

This is why, one has to think about constructing a pdf using the computed values for δ :

$$pdf(\vec{r}) \propto 1 + \delta(\vec{r}) \quad [39]$$

Also ensuring the non-negativity of this quantity is a problem : some values of δ are under -1. In the meantime, a density function cannot be negative and this excludes this case as a physical one. Three methods are thus possible :

- Using a log-normal pdf (14):

$$pdf(\vec{r}) \propto \exp(\delta(\vec{r})) \quad [40]$$

- Dilating the field :

$$pdf(\vec{r}) = 1 + c\delta(\vec{r}) \quad [41]$$

with $c \leq -\frac{1}{\delta_{min}}$.

- Truncating the field :

$$pdf(\vec{r}) \propto 1 + \mathbb{I}_{\delta(\vec{r})>-1}\delta(\vec{r}) \quad [42]$$

In this work, second and third methods were used and mostly the last.

Once the pdf is computed in the box, the rejection method (15) allows to randomly draw the desired number of objects either discretely on centers of pixel or semi-continuously on continuous space values thanks to a 3D-interpolated pdf. Pdf computation and both continuous and semi-continuous methods are implemented in the *cataloger.py* code. An option permits visualization of the drawing as showed in figure 6.

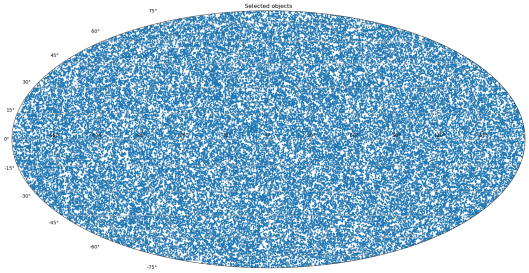

Fig. 6. Mollweid map of angular dispersion of 60000 objects selected at different redshifts in a 256^3 box.

In order to study the results at a given redshift and to be more realistic, *cataloger.py*, also includes the possiblity to change the pdf to :

$$pdf(\vec{r}) \propto 1 + \mathbb{I}_{\delta(\vec{r})>-1} D_+(z) b(z, M) \delta(\vec{r}) \quad [43]$$

with for instance the bias (16) set as :

$$b(z, M) = 1 + \frac{(\frac{\delta_c(z)}{D_+(z)\sigma(M,0)})^2 - 1)}{\frac{\delta_c(z)}{D_+(z)}} \quad [44]$$

Unified draw from pdf The previous instructions do not allow to generate random objects with a mass. This last should follows (as a first approximation) the Press-Schechter distribution (see eq. 9). This is how we elaborate a method to generate a catalog following both the space correlation function values and the Press-Schechter predictions :

- Cut the box into iso-redshift shells $[z - dz/2, z + dz/2]$
- For each shell :
  - Compute the pdf on the associated pixels.
  - Draw counts following the Press-Schechter Poisson law for chosen bins of mass (here from $1.6e14$ to $1e16 M_\odot$ with 0.01 log-bins to get around 200000 objets).
  - Following the pdf and thanks to the rejection method, draw as much positions as there are objects (any mass).
  - (Discrete case) Assign all the counts to the positions with the following rule : the greater δ is, the more massive is the object. (Continuous case, not implemented) The same with an interpolated pdf (from the discrete box).

This is typically what is done in the file *DoubleCataloger.py* also with a discrete or semi-continuous approach. Note that it also produces a randomly distributed catalog, useful to compute the correlation thanks to Landy and Szaslay (see eq. 16). In fact, conversely to a simple draw without mass drawing, the number of draw per shell is not constant and that is why the random catalog has to be adapted to keep the same counts density per shell.

Also, the counts in each bins are always saved in a second file to use quickly the catalog without having to count again.

Correlated counts draw One another method allows to build correlated counts only. It takes advantage of the growth factor and bias modification of the box density field to dilate the expected number of counts before drawing the counts with the Poisson law :

$$\overline{N}'(z, M) = (1 + D_+(z) b(z, M) < \delta >_z) \overline{N}(z, M) \qquad [45]$$

This is also included in *DoubleCataloger.py* as *GenerateCorrelatedCounts* method.

D. Results. At first, the generation of a catalog of counts is not a problem as testifies the successful refinement showed in figure 3 (see *data.py*). In fact, once the different quantities useful for the computing of theoretical mean binned counts (see *Cosmological_tools.py*) are implemented, a Poisson law draw allows to get one. We just notice a slight reduction of contours' size compared to Planck 2015 collaboration (5). Also, the Gaussian likelihood gave the same results.

At the stage of the box, the correlation varies depending on the box draw (5 is one realization). With 20 boxes, we are able to provide the mean correlation and the standard deviations on figure 7.

The correlation seems a bit different and even the use of an optimized spherical convolution function (against pixels' effect) did not helped to bring closer the theoretical curve. We let this slight difference aside for the moment, acknowledging it would require some more in depth checks.

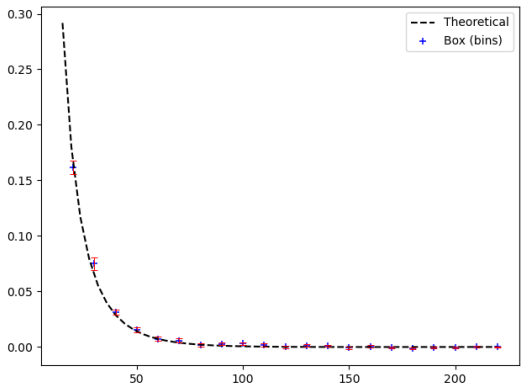

Fig. 7. Mean space correlation function from 20 box of size 256^3 with 20 bins.

The next step was to generate 20 catalogs from the 20 boxes. Such a Monte-Carlo routine is under comments at the end of *cataloger.py*.

Then the Landy and Szaslay estimator gives the correlation for each catalog (see *dataCorrelation.py*, using 3 times more points in the random catalog). With 20 realizations, we are able to provide the mean points and their uncertainty on figure 8, for the case of a discrete catalog. A new bias is revealed in this figure, but different from the previous one and of greater importance (to be compared with 7). The bias is even more revealed here when inferring the model on this data, giving around 0.04 errors on cosmological parameters while there were not such errors as inputs. One has also to consider that we did not find any geometrical anomaly in the correlation by using pixel-centered positions for the generation of the catalog since the Landy and Szaslay estimator was used with the same geometry for the random catalog.

The cosmological constraints can be visualized in two different ways. The first method is simply about plotting contours of the likelihood between the mean data and the theory. However, to avoid the bias on this figure, we replace the data by the theoretical curve but keep the standard deviations from the data. Results are shown in figure 9. The second way is an approximation of a Monte Carlo Markov Chain method as we group the dispersion of the 20 refined couples in terms of contours. Two sub-methods were used : a linear interpolation of the density of points and the consideration of convex hulls for given levels of confidence. Results are shown in figure 10.

For such low amounts of points, we see there the limit of the interpolation in the first method even if it is a smoothing method compared to convex hulls. Furthermore, the size of the convex hulls are in good agreement with the ellipses of figure 9. We note that counts contours (figure 3) are smaller than all those contours but also almost perpendicular, which would help to break degeneracies.

We also must precise that the choice of 200000 was not arbitrary since error bars were too consequent for only 60000 objects for instance. This paves the way to the importance of upcoming Euclid survey with such amounts of data.

Today we don't know the origin of the bias in the draw of the catalog. A first possibility was still about geometrical effects since we chose objects at the center of the pixels. For

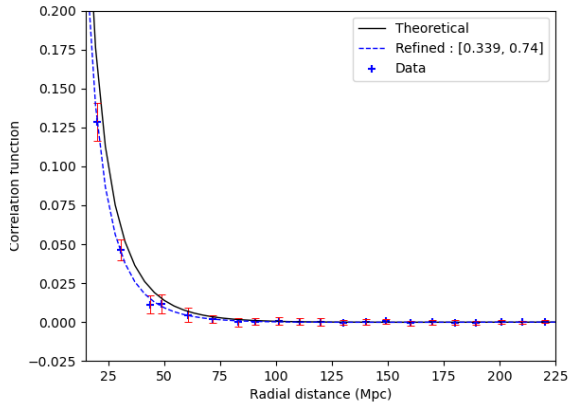

Fig. 8. Mean correlation data from $20\ 256^3$ discrete catalogs (from 20 different boxes) and its associated refined model versus the theoretical correlation ($\Omega_m = 0.3, \sigma_8 = 0.8$). Landy and Szaslay estimator implemented with *nbodykit* in *dataCorrelation.py*. Random catalogs had also pixels' centers as available positions.

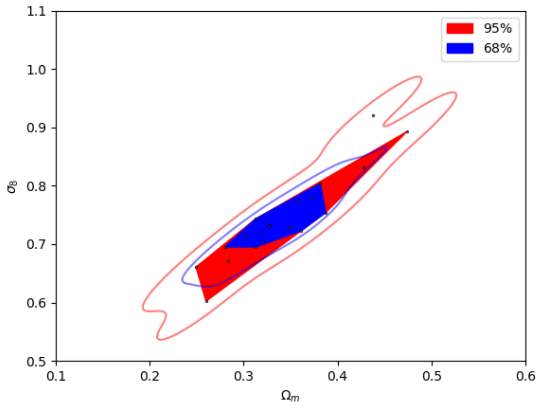

Fig. 10. Confidence contours of 20 refined values for (Ω, σ_8) box of size 256^3 with 20 bins. Method 1 (lines) : linear interpolation. Method 2 (areas) : Convex hulls of the points (19 points out of 20 corresponding to 95%).

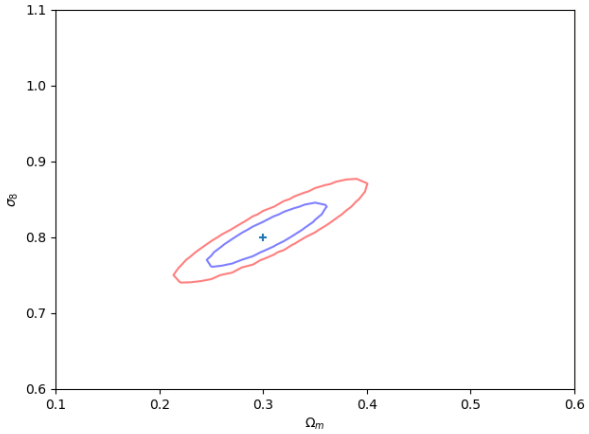

Fig. 9. Theoretical confidence contour brought by the refinement of the theoretical correlation with the standard deviations from 8.

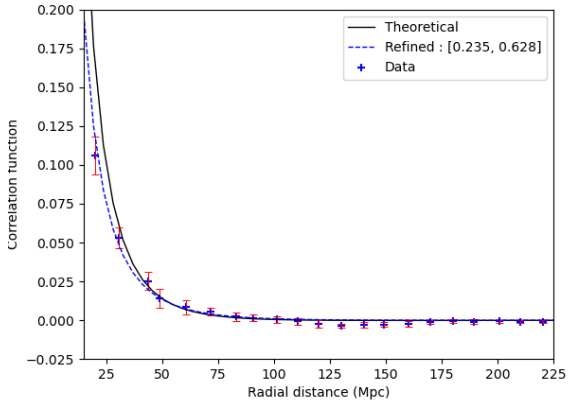

Fig. 11. Mean correlation data from 20×20^3 continuous catalogs (from 20 different boxes) and its associated refined model versus the theoretical correlation ($\Omega_m = 0.3, \sigma_8 = 0.8$). Landy and Szaslay estimator implemented in *dataCorrelation.py*.

the moment *cataloger.py* includes a 'continuous' mode where positions are generated with a continuous rejection method thanks to the 3D interpolation of the discrete pdf. However, for such a draw, 200000 were not sufficient to find the correlation with Landy and Szaslay estimator. In fact, the continuous method gives good results in the same framework as the continuous calculus of the correlation in the box, ie in a sub-box or approximately $1e4$ evaluations per 20^3 pixels, requiring round $1e7$ objects in the catalog. Figure 11 gives the resulting correlation for 20 boxes with the associated error bars. The continuous draw does not seem to delete completely the bias. Moreover, it concedes a too great number of objects compared to reality.

With this second try, we expect the bias to be coming from the choice of the pdf. Choosing eq. 50 or 51 may be more adapted for the drawing of objects starting from such a box. We tested eq. 51 without a lot of success but it still needs attention for a future try.

In the following, we set aside this bias difficulty to go on the double catalog matter, consciously knowing that any bias brought previously in counts or correlation simulations will be reproduced in the unified catalog.

The natural algorithm described in C.2 for a "unified draw" has been successfully implemented, at least for both eq. 43 and 45. In fact, the use of rejection method in each shell induces the right input correlation (8) if we pay attention to the conservation of counts per redshift shell in the random catalog as explained in C.2.

This confirms the coherence between *cataloger.py* and *DoubleCataloger.py* correlations. For the counts, the coherence with *data.py* was also checked since the same lines of codes were used.

With such a catalog with correlated positions and masses, it is possible to compute the global covariance matrix (eq. 22) thanks to 20 double catalogs from 20 boxes. The results are showed on figures 12 and 13 in the case of eq. 43.

Many lessons come out of this figure. As it was stated in the past (9), off-diagonal counts terms are not systematically negligible in front of the diagonal ones : here they reach near a half of correlation. This is the same conclusion for off-diagonal correlation terms and also for $\Sigma_{N,\xi}$ terms which are for the first time measured on this figure. All these elements constitute

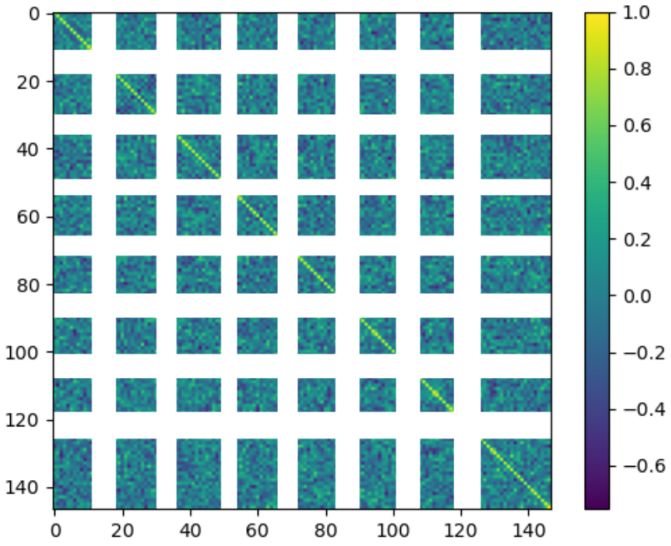

Fig. 12. Counts-correlation correlation matrix (correlation associated to the covariance Σ eq. 22) of 10 catalogs from 10 different boxes. Bins from 1 to 126 : full-sky counts bins from redshifts 0 to 0.7 (7 bins with 0.1 steps) each one with sub-bins for masses from $1.6e14 M_\odot$ to $1e16 M_\odot$ (18 bins with 0.1 log steps). Bins from 127 to 148 : 21 correlation bins increasing with radius from 20 to 220 (same as fig. 11). Blank correlations are due to absence of counts at high-redshifts.

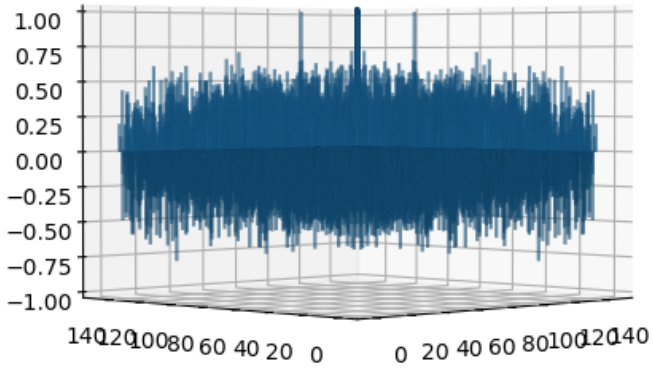

Fig. 13. Counts-correlation correlation matrix (correlation associated to the covariance Σ eq. 22) of 10 catalogs from 10 different boxes. Bins from 1 to 126 : full-sky counts bins from redshifts 0 to 0.7 (7 bins with 0.1 steps) each one with sub-bins for masses from $1.6e14 M_\odot$ to $1e16 M_\odot$ (18 bins with 0.1 log steps). Bins from 127 to 148 : 21 correlation bins increasing with radius from 20 to 220 (same as fig. 11). Blank correlations are due to absence of counts at high-redshifts.

a call for more consideration on correlations between counts and space correlations.

To end with the main results obtained in simulation and because we have the opportunity to refine two observables on the same catalog, we evoke here a first attempt to refine both space correlation bins and counts bins with a simple (Σ)-diagonal Gaussian model. Paradoxically thanking the bias we had on correlation, the conclusion is quite obvious : the constraints are still good and refer to the constraints we already had with counts only (see figure 3) despite the bad trend of correlation data points which should shift the constraints from the $(0.3, 0.8)$ input point. This shows that our standard deviations are too high for the space correlation function and do not allow to gain on cosmological constraints with this configuration on the simulations.

4. The search for a unified theory

A. Press-Schechter postulates' modification. Press-Schechter (2) formalism starts from the hypothesis of a δ-field of independent Gaussian random variables and we saw previously that it was not compatible with non-null correlations. To go further than a unified likelihood, we propose here to unify correlation function and counts from the beginning in the construction of the counts prediction.

The first step is to postulate the δ-field (at a given time and then any time deduced thanks to linear dilation) as a realization of a (continuous or discrete at this stage) random Gaussian field (or stochastic Gaussian process) where each point i is correlated to all of them. Mathematics allow then to describe each finite random n-vector $(\delta_i)_{1\geq i\geq n}$ as a multivariate Gaussian variable :

$$(\delta_i)_{1\geq i\geq n} \sim \mathcal{N}(0, \Sigma) \qquad [46]$$

The real postulate comes then in the construction of Σ. We set now $\boldsymbol{\delta}_{ini} = (\delta_i)_{1\geq i\geq n}$ as the initial field (at z_{ini} for instance) and let be V a volume containing all the considered points. As we already mentionned in B.1 and especially in eq. 26, ξ_R appears to be a good candidate for covariance under the right assumptions. We suppose it is the case and write :

$$\Sigma_{ij} = \xi_R(\vec{x}_i - \vec{x}_j, z_{ini}) = \xi_R(\vec{x}_i - \vec{x}_j) \qquad [47]$$

This writing allows to find back diagonal terms of the Press-Schechter distribution since $\Sigma_{ii} = \xi_R(\vec{0}) = \sigma^2(R)$, see eq. 4.

The probability density function at a given $\boldsymbol{\delta}$ vector point writes :

$$\mathbb{P}(\boldsymbol{\delta}_{ini} = \boldsymbol{\delta}) = \frac{1}{\sqrt{2\pi}^n |\Sigma|^{\frac{1}{2}}} e^{-\boldsymbol{\delta}^T \Sigma^{-1} \boldsymbol{\delta}} \qquad [48]$$

This works then takes advantage of the property of a covariance matrix : it is always semi-definite positive and thus admits a Cholesky decomposition : $\Sigma = AA^T$ with A a lower triangular matrix, the diagonal of which is non-negative. So : $(A^{-1}\boldsymbol{\delta})^T (A^{-1}\boldsymbol{\delta})$ and then the variable change $\boldsymbol{x} = A^{-1}\boldsymbol{\delta}$ is such that :

$$\mathbb{P}(\boldsymbol{x}_{ini} = \boldsymbol{x}) = \frac{1}{\sqrt{2\pi}^n} e^{-\boldsymbol{x}^T \boldsymbol{x}} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-x_i^2} \qquad [49]$$

In fact, we just proved the common result that each centered Gaussian vector admits a matricial transformation allowing to get a vector of independent Gaussian variables.

B. Predictions. An application of the Press-Schechter principle is possible then : we identify the virialized objects at the places where the fluctuations are over the δ_c threshold at a given redshift z. This threshold is the same for any point and can be brought to z_{ini} thanks to the linear growth factor as in eq. 3. We denote $\boldsymbol{x_0} = A^{-1}(\delta(z_{ini}))$.

The probability for k points to be virialized objects over R scale and not $k-n$ given a γ permutation is :

$$\begin{aligned}&\mathbb{F}_{k,\gamma}(>R,z) = \\ &\mathbb{P}(\delta_{ini,\gamma(1)} > \delta(z_{ini}), ..., \delta_{ini,\gamma(k)} > \delta(z_{ini}), \\ &\delta_{ini,\gamma(k+1)} \leq \delta(z_{ini}), ..., \delta_{ini,\gamma(n)} \leq \delta(z_{ini})) \\ &= \int_{x_{0,\gamma(1)}}^{\infty} ... \int_{x_{0,\gamma(k)}}^{\infty} \int_{-\infty}^{x_{0,\gamma(k+1)}} ... \int_{-\infty}^{x_{0,\gamma(n)}} \frac{1}{\sqrt{2\pi}^n} e^{-\boldsymbol{x}^T \boldsymbol{x}} d^n\boldsymbol{x} \\ &= \prod_{i=1}^{k} \frac{1}{2}(1 - erf(x_{0,\gamma(i)})) \prod_{i=k+1}^{n} \frac{1}{2} erf(x_{0,\gamma(i)})\end{aligned} \tag{50}$$

As a good approximation, taking δ_{ini} as a grid with R-scale pixels allows to predict the expected number of virialized object of scale over R (equivalent to $> M$). In fact, taking into account the permutations :

$$\overline{N}(>M,z) = \sum_{k=1}^{n} k \sum_{\gamma} \mathbb{F}_{k,\gamma}(>M,z) \tag{51}$$

With a theoretical Σ computed for a given grid, one is able to compute $A(M)$ (no simple explicit formula) and deduce the previous number. However, there is no way to deduce the distribution depending on mass since it depends on grid's choice. And I am not even sure that 51 is correct since the grid is only for near R scales.

This is why we must find the virializing probability in $[M, M+dM]$ to use the equivalent of eq. 6. One way to find this probability is to compute the proportion of virialized objects ($> M$) and derive according to M:

$$\begin{aligned} q(M) = & \frac{\partial}{\partial M} \sum_{k=1}^{n} \frac{k}{n} \sum_{\gamma} \mathbb{F}_{k,\gamma}(>M,z) \\ = & \sum_{k=1}^{n} \frac{k}{n} \sum_{\gamma} \frac{\partial}{\partial M} \mathbb{F}_{k,\gamma}(>M,z) \end{aligned} \tag{52}$$

with the derivation :

$$\begin{aligned}&\frac{\partial}{\partial M}\mathbb{F}_{k,\gamma}(>M,z) \\ &= \frac{1}{2^n} \prod_{i=1}^{k}(1 - erf(x_{0,\gamma(i)})) \\ &\times \sum_{j=k+1}^{n} \frac{\partial x_{0,\gamma(j)}}{\partial M} e^{-x_{0,\gamma(j)}^2} \prod_{i=1,\neq j}^{k} erf(x_{0,\gamma(i)}) \\ &- \frac{1}{2^n} \prod_{i=k+1}^{n} erf(x_{0,\gamma(i)}) \\ &\times \sum_{j=1}^{k} \frac{\partial x_{0,\gamma(j)}}{\partial M} e^{-x_{0,\gamma(j)}^2} \prod_{i=1,\neq j}^{k}(1 - erf(x_{0,\gamma(i)}))\end{aligned} \tag{53}$$

where :

$$\frac{\partial x_{0,\gamma(j)}}{\partial M} = \delta(z_{ini}) \sum_{j=1}^{n} \frac{\partial (A^{-1})_{i,\gamma(j)}}{\partial M} = \frac{\delta_c(z) D_+(z_i)}{D_+(z)} \sum_{j=1}^{n} \frac{\partial (A^{-1})_{i,\gamma(j)}}{\partial M} \tag{54}$$

$\frac{dn}{dM}$ follows then from 6. Note that we find back Press-Schechter distribution by computing $\mathbb{P}(\delta_{ini,i} > \delta(z_{ini}))$ with a diagonal (and so isotropic) Σ.

This new method should give a better way to estimate to virializing probability for a given correlated grid.

5. Last discussions and conclusions

In this work, many frameworks have been explored.

As a first incursion, the discovery of the theory lead me to notice many similarities in the formalism of the two studied observables, despite the use by very different communities in practice.

As the main approach of the physicists of the domain is about the search of cosmological constraints, we managed to put in practice both their refinement. This was the last step with isolated theories.

If the ultimate goal of the subject was about the unification of those two formalisms, nothing could be done without a unification of catalogs.

This is how came the idea of a simulated box of universe. For such a short period, the little errors we still have do not hide the success of the approach since we were successful in the production of an artificial catalog and leading to the first measure of a Counts - Space Correlation correlation matrix. The remaining bias on space correlation function will need a bit more attention, before applying last updates of the code with growth factor and cluster bias inclusions to be more realistic. In fact, the git repository includes a bit more files than used and tested ones. I encourage the next person to exploit elements of my code. Moreover, as a proposition and now that it is almost mastered, maybe we should use an already existing software (if existing) if we don't loose the flexibility and control we have on the current simulator.

As an irresistible desire of mine, the goal was also to explore new theoretical possibilities in the unification of the likelihood through both the study of a generalized covariance and the study of a unified observable.

About that, we warn the reader on the approaches which were developed here. They still need review and applications : for instance in section 3 the definition of the correlation is still a debate for me : does it need to be a random variable and what is the real validity of the ergodic principle ? Also, in the multivariate Press-Schechter proposition (section 4), many questions and developments will need to be put on a sheet of paper.

This remark is valuable for many results of this paper : a poor maturity on the subject is expected for a 2-month internship and that is why I encourage any person reading this paper to contact me for any new element of response or questioning.

6. References

1 E Artis, Ph.D. thesis (Paris-Saclay University) (2021).
2 WH Press, P Schechter, Formation of Galaxies and Clusters of Galaxies by Self-Similar Gravitational Condensation. **187**, 425–438 (1974).
3 JM Bardeen, JR Bond, N Kaiser, AS Szalay, The Statistics of Peaks of Gaussian Random Fields. **304**, 15 (1986).
4 G Holder, Z Haiman, JJ Mohr, Constraints on Ω_m, Ω_Λ, and σ_8 from Galaxy Cluster Redshift Distributions. **560**, L111–L114 (2001).
5 Planck Collaboration, et al., Planck 2015 results. XXIV. Cosmology from Sunyaev-Zeldovich cluster counts. **594**, A24 (2016).
6 DI Inouye, E Yang, GI Allen, P Ravikumar, A review of multivariate distributions for count data derived from the poisson distribution. *WIREs Comput. Stat.* **9**, e1398 (2017).
7 K Kawamura, The structure of multivariate Poisson distribution. *Kodai Math. J.* **2**, 337 – 345 (1979).
8 A Morin, A propos de la loi de poisson multivariée. *J. de la société française de statistique* **134**, 3–13 (1993).
9 W Hu, AV Kravtsov, Sample Variance Considerations for Cluster Surveys. **584**, 702–715 (2003).
10 H Dole, *Cours de cosmologie moderne.* (year?).
11 SD Landy, AS Szalay, Bias and Variance of Angular Correlation Functions. **412**, 64 (1993).
12 JN Grieb, AG Sánchez, S Salazar-Albornoz, C Dalla Vecchia, Gaussian covariance matrices for anisotropic galaxy clustering measurements. *Mon. Notices Royal Astron. Soc.* **457**, 1577–1592 (2016).
13 HT Yura, SG Hanson, Digital simulation of an arbitrary stationary stochastic process by spectral representation. *J. Opt. Soc. Am. A* **28**, 675–685 (2011).
14 P Coles, B Jones, A lognormal model for the cosmological mass distribution. **248**, 1–13 (1991).
15 WR Gilks, NG Best, KKC Tan, Adaptive rejection metropolis sampling within gibbs sampling. *J. Royal Stat. Soc. Ser. C (Applied Stat.* **44**, 455–472 (1995).
16 HJ Mo, SDM White, An analytic model for the spatial clustering of dark matter haloes. *Mon. Notices Royal Astron. Soc.* **282**, 347–361 (1996).

I want to sincerely thank my supervisor Jean-Baptiste Melin for this experience, his help and advices and also people at the lab who remained attentive for my work and personal development : Corentin Ravoux, Edmond Chaussidon, Arnaud De Mattia ...