Draft version October 13, 2023
Typeset using LaTeX **manuscript** style in AASTeX631

Angular correlation function of quasars*

Astrid Benamou[1] and Roya Mohayaee[2]

[1] *Université Dauphine, Paris 75016*

[2] *IAP, 98 bis bld Arago, Paris 75014*

ABSTRACT

We show how to best characterise astronomical datasets. Correlation function is often used to characterise the distribution functions. We first use "mock" datsets, that we know the distribution of. This allows us to see if the correlation function is a good estimator of the statistical properties of a distribution. We study various artificial shortcomings, e.g. the size of the sample, the box size, the shot noise etc. We then run our estimator on the real observation data from high redshift quasars. Study angular correlation functions etc.

We study the position of quasars given by spherical coordinates aroung the observation point. To evaluate the distribution of their locations the angular correlation function is an efficient tool that we will be using in this article.

Keywords: Astronomy, cosmology, statistical physics, data analysis

1. INTRODUCTION

The observational data in astronomy are usually discrete: *i.e.* at best we observe galaxies and their positions (θ, ϕ, z in spherical coordinates usually). We often idealise and say consider our data as particles in a box, which we need to analyse and understand. The dataset is huge, so often all we can do is to characterise the data statistically. For instance, we find the distribution function, or find the

* Stage Astrid June, July, August , 2021

two point, three point and higher order correlation functions. Otherwise, often people use Fourier analysis, go to Fourier space and get the power spectrum. There are many "spurious" effects which we need deal with in order to get meaningful results.

Before analysing real observational data, we want to grasp various difficulties by using idealised simulations. Hence, as first exercise, we choose a distribution (here, uniform, Gaussian and log-normal) and then make a realisation of that distribution by randomly picking particles from that distribution. We then pretend that we did not know our initial distribution (which is the case when one deals with real data) and try to see if by using correlation function analysis we can find our input distribution.

We wish to show if we pick particles randomly from a distribution function (i.e. go from a continuous to a discrete representation) then we can (or cannot) recover our original distribution function or all the information about it. Due to discretisation we are also effected by Poisson (shot) noise, box size and granularity (i.e; particle separation). We wish to first study these effects to see how to properly do the analysis to get maximum information about the original distribution.

In the first part, we will simulation particles in a box following a known distribution and calculate the associated pair correlation function. We will also try to find the distribution from the correlation function calculated on the database in the xz0 file.

Then, we will transform the three Cartesian coordinates data to a two spherical coordinates which is closer to the real data we have in the catalogue. This is dued to the fact that often we are not able to obtain the red shift associated to each entity. We will therefore have to calculate the Angular correlation function based only on the two coordinates: the angles ϕ and θ.

Finally, we will try to analyse the catalogue of quasars by calculating the angular pair correlation between the points given.

Our main motivation is that the data is so dense we can only analyse it using statistical tools. Furthermore, the pair correlation function is a good way to analyse the distribution of the sample as it characterizes it very well and is easy to calculate.

(a) 3D representation of a sample of 1000000 particules following a uniform distribution in the box

(b) Density function of a uniform on [0,1]

Figure 1: Uniform distribution on [0,1]

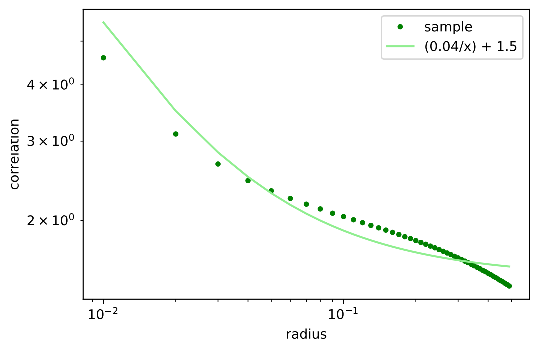


Figure 2: Correlation function for a uniform distribution

2. SIMULATION OF PARTICLES IN A BOX DRAWN FROM A PARTICULAR DISTRIBUTION AND THEIR CORRELATION

2.1. *Uniform distribution : density and its correlation function*

You can calculate these functions for different number of particles in order to study the noise in your sample. Try to increase the number of points to about 1000000 to 2000000 (so that you are prepared to work with the real catalogue of quasars later on).

To simulate particules following a uniform distribution in a box with sides of lenght 1, we simulate the three coordinates of each particule (x,y,z) following a uniform distribution on [0,1].

Density—If X follows a uniform distribution on [0,1], its density is given by :

$$f(x) = \mathbb{1}_{[0,1]}(x)$$

Hint: You will see that the correlation function you obtain for a uniform distribution is not zero, which is what we expect. The reason is the small fluctuation, shot noise, box size etc... So when we analysing any other distribution, one needs divide by the this uniform correlation function in some sense to normalise and to take away the effect of noise. The histogram approach is good but do not forget to normalise. Correlation function for random points is "defined" to be zero.

This is exactly the reason why it is defined as the excess number of pairs of points separated by a distance r, relative to what it would be if the points were randomly distributed. The definition is:

$$\xi(r) = \frac{DD(r)\Delta r}{RR(r)\Delta r} - 1 \tag{1}$$

where $DD(r)\Delta r$ is the number of pairs of data points separated by a distance $r \pm \Delta r$ and $RR(r)\Delta r$ is the corresponding number of pairs if the point process is Random. The correlation function is thus defined to be zero for a set of random points. There are different estimators, above is the simplest and rather adequate method.

2.2. *Gaussian distribution and its density and correlation function*

To simulate particles following a Gaussian distribution, we need to make sure the values of the simulation fall between 0 and 1 for them to correspond to the coordinates of the particles in the box.

To do so we take a Gaussian of mean $\mu = 0.5$ and a small variance, for example, $\sigma^2 = 0.2$. We then truncate the law to keep only values between 0 and 1.

Density—The density of a gaussian with mean μ and variance σ^2 is :

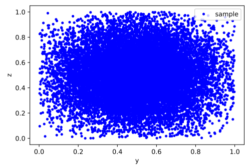
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2)$$

Denote the cumulative distribution of the same gaussian by F.

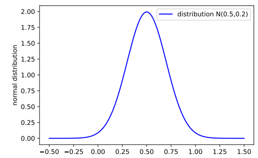
And the density of a gaussian with the same parameters truncated to [a,b] is :

$$f_{trunc}(x, a, b) = \frac{f(x)}{F(b) - F(a)}$$

Here we take a = 0 and b = 1.



(a) 3D representation of sample



(b) Density function

Figure 3: Gaussian distribution with parameters $\mu = 0.5$ and $\sigma = 0.2$

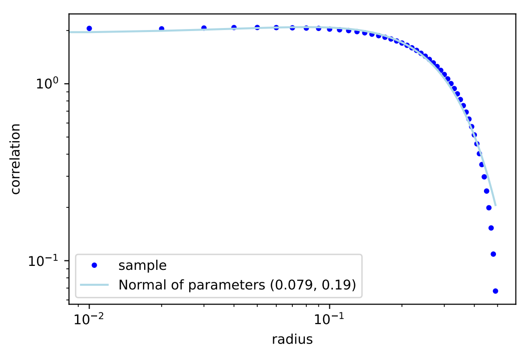


Figure 4: Correlation function for a sample of 100000 particules following a normal distribution

We can fit the pair correlation by the density of a gaussian truncated to [0,0.5] of parameters $\mu = 0.079$ and $\sigma = 0.19$ using the function optimize.curve_fit from scipy.

Show analytically that for a Gaussian distribution, two point correlation function contains all the information about the distribution.

$$
\begin{aligned}
< x > &= \int_{-\infty}^{\infty} dx\, xP(x) \\
&= \tfrac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty}\ xexp(-\tfrac{(x-\mu)^2}{2\sigma^2})\, dx \\
&= \tfrac{\sqrt{2}\sigma}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sqrt{2}\sigma t+\mu)\, exp(-t^2)\, dt \\
&= \tfrac{1}{\sqrt{\pi}}(\sqrt{2}\sigma \int_{-\infty}^{\infty} texp(-t^2)\, dt + \mu \int_{-\infty}^{\infty} exp(-t^2)\, dt) \\
&= \tfrac{1}{\pi}(\sqrt{2}\sigma\, [-\tfrac{1}{2}exp(-t^2)]_{-\infty}^{\infty} + \mu\sqrt{\pi}) \\
&= \tfrac{\mu\sqrt{\pi}}{\sqrt{\pi}} \\
&= \mu
\end{aligned}
\tag{2}
$$

$$
\begin{aligned}
< x^2 > &= \int_{-\infty}^{\infty} dx\, x^2 P(x) \\
&= \tfrac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty}\ x^2 exp(-\tfrac{(x-\mu)^2}{2\sigma^2})\, dx \\
&= \tfrac{\sqrt{2}\sigma}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sqrt{2}\sigma t+\mu)^2\, exp(-t^2)\, dt \\
&= \tfrac{1}{\sqrt{\pi}}(2\sigma^2 \int_{-\infty}^{\infty}\ t^2\, exp(-t^2)dt + 2\sqrt{2}\sigma\mu[-\tfrac{1}{2}exp(-t^2)]_{-\infty}^{\infty} + \mu^2\sqrt{\pi}) \\
&= \tfrac{1}{\sqrt{\pi}}(2\sigma^2 \int_{-\infty}^{\infty}\ t^2 exp(-t^2)dt + 2\sqrt{2}\sigma\mu * 0) + \mu^2 \\
&= \tfrac{2\sigma^2}{\sqrt{\pi}}\ \int_{-\infty}^{\infty} t^2\, exp(-t^2)\, dt \\
&= \tfrac{2\sigma^2}{\sqrt{\pi}}\, ([-\tfrac{t}{2}exp(-t^2)]_{-\infty}^{\infty} + \tfrac{1}{2} \int_{-\infty}^{\infty}\ exp(-t^2)\, dt \\
&= \tfrac{2\sigma^2}{\sqrt{\pi}}\ \tfrac{1}{2} \int_{-\infty}^{\infty}\ exp(-t^2)\, dt \\
&= \tfrac{2\sigma^2\sqrt{\pi}}{2\sqrt{\pi}} \\
&= \sigma^2
\end{aligned}
\tag{3}
$$

For the next equations we use an integration by part and we take mu = 0:

$$\begin{aligned}
< x^3 > &= \int_{-\infty}^{\infty} dx\, x^3 P(x) \\
&= \tfrac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x^3 exp(-\tfrac{x^2}{2\sigma^2})\, dx \\
&= \tfrac{-\sigma^2}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 \tfrac{-x}{\sigma^2} exp(-\tfrac{x^2}{2\sigma^2})\, dx \\
&= \tfrac{-\sigma}{\sqrt{2\pi}}([x^2 exp(-\tfrac{x^2}{2\sigma^2})] - \int_{-\infty}^{\infty} 2x exp(-\tfrac{x^2}{2\sigma^2})\, dx) \\
&= \tfrac{-\sqrt{2}\sigma}{2} < x > \\
&= 0 if \mu = 0
\end{aligned} \tag{4}$$

$$\begin{aligned}
< x^4 > &= \int_{-\infty}^{\infty} dx\, x^4 P(x) \\
&= \tfrac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x^4 e^{\frac{x^2}{2\sigma^2}}\, dx \\
&= -\tfrac{\sigma^2}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x^3 \tfrac{-x}{2\sigma^2} e^{\frac{x^2}{2\sigma^2}}\, dx \\
&= -\tfrac{\sigma}{\sqrt{2\sigma}}([x^3 e^{-\frac{x^2}{2\sigma^2}}] - \int_{-\infty}^{\infty} 3x^2 e^{-\frac{x^2}{2\sigma^2}}\, dx) \\
&= \tfrac{\sigma}{\sqrt{2\sigma}} 3 \int_{-\infty}^{\infty} x^2 e^{-\frac{x^2}{2\sigma^2}}\, dx \\
&= \tfrac{3\sigma}{\sqrt{2\sigma}} < x^2 >
\end{aligned} \tag{5}$$

where $P(x)$ is your distribution function.

2.3. *Lognormal distribution and its density correlation function*

A log-normal distribution of parameters μ and σ has the following density :

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} exp(-\frac{(lnx-\mu)^2}{2\sigma^2}) \tag{6}$$

and we use the same formula as before to obtain its density truncated to [0,1]

(a) Representation of a slice of the sample (b) Density function

Figure 5: Log-normal distribution

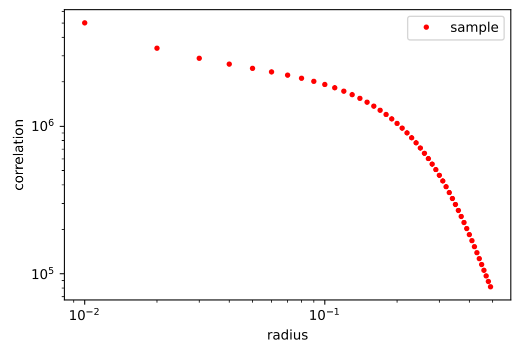
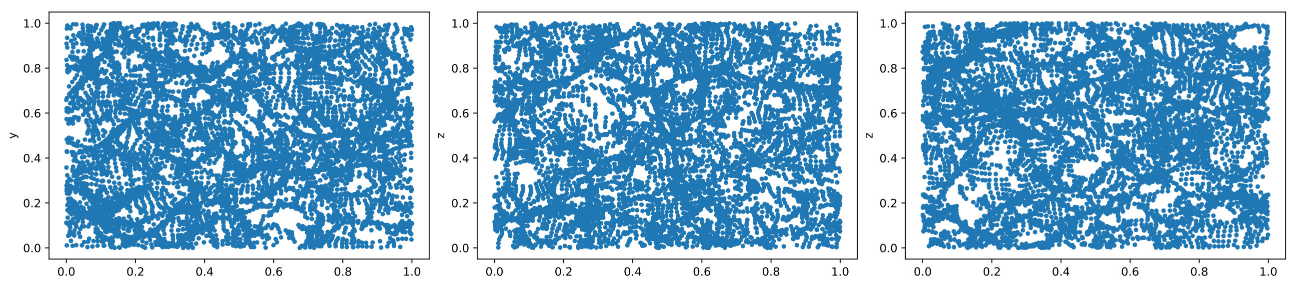


Figure 6: Correlation function for a sample of log-normal distributed particles.



(a) X-Y for Z $\in [0, 0.01]$ (b) Y-Z plot for X $\in [0, 0.01]$ (c) X-Z plot for Y $\in [0, 0.01]$

Figure 7: slices of the data that shows the data is not uniform. 3D representation is illegible.

2.4. *First set of data "xz0"*

By analysing the other slices of data we see that they look similar to this distribution of points.

I started by calculating the pair-correlation function on a sample of size 180 000 the data randomly picked from the data base and obtained the following correlation function :

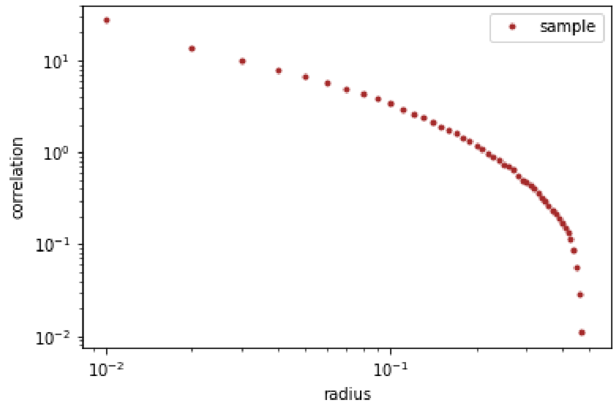


Figure 8: Correlation Function on a sample of the first 200 000 particles using the Landy Szalay estimator in log scale.

This correlation could correspond to a lognormal distribution according to our previous study.

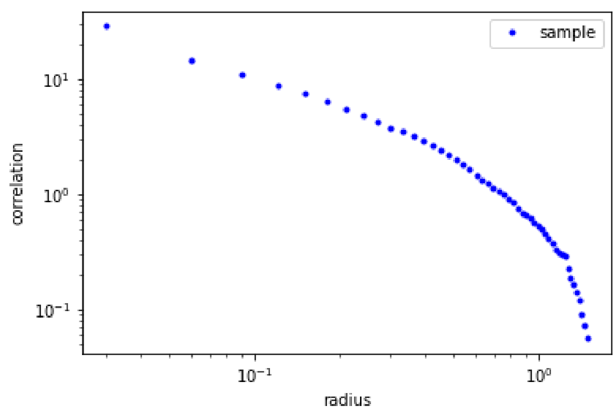


Figure 9: Correlation Function, for data from a real cosmological simulation at $z = 0$, (corresponding to a lognormal distribution) on a sample of 200 000 particles with changed x-axis to r/a with a = 0.3719 the mean interparticle seperation.

3. PARTICLES IN A SPHERE CENTRED ON THE BOX

In this section, we do a simple exercise and just go from Cartesian to spherical coordinates and find the one and two points correlation function for the same distributions.(This is a preparation section for the next section which is to go and project all particles on the sky and produce a "mock"

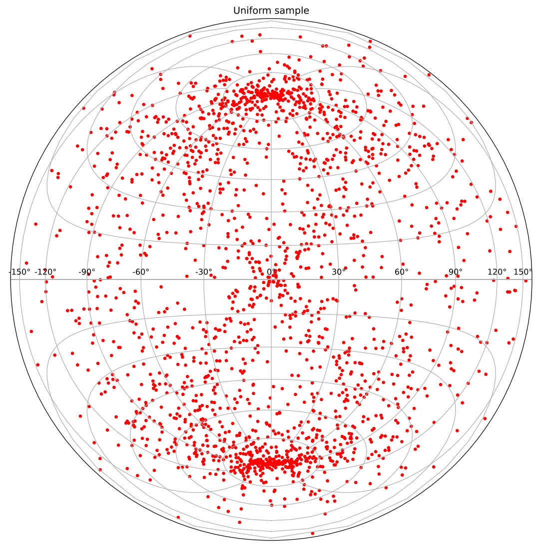


Figure 10: 3D representation of uniform sample on a sphere

sky catalogue which would resemble our real quasar data. This time cut a sphere inside your box and calculate the same quantities inside.

In this section we want to simulate the coordinates of particles in a sphere. Therefore these coordinates are :

- The radius with values between 0 and 1.
- The 2 angles Theta between 0 and 2π and Phi with values between $-\frac{\pi}{2}/$ and $\frac{\pi}{2}/$.

3.1. *Uniform distribution : density and its correlation function*

You can calculate these functions for different number of particles in order to study the noise in your sample.

We will demonstrate analytically why a uniform sample in a box is not uniformly distributed on a sphere :

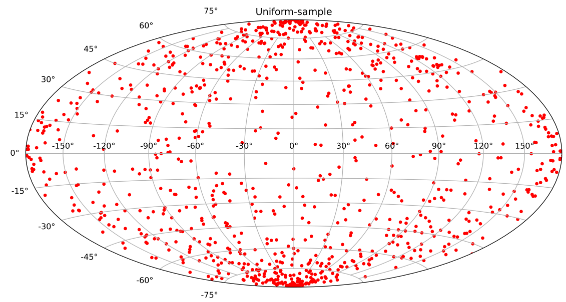


Figure 11: Flat representation of uniform sample

The transformation between Cartesian coordinates and spherical coordinates is given by the following formulas :

$$\begin{aligned} x &= \rho sin(\phi)cos(\theta) \\ y &= \rho sin(\phi)sin(\theta) \\ z &= \rho cos(\phi) \end{aligned} \tag{7}$$

If we calculate the jacobian of this change of variables we obtain :

$$Jacobian = \begin{vmatrix} \frac{\delta x}{\rho} & \frac{\delta x}{\phi} & \frac{\delta x}{\theta} \\ \frac{\delta y}{\rho} & \frac{\delta y}{\phi} & \frac{\delta y}{\theta} \\ \frac{\delta z}{\rho} & \frac{\delta z}{\phi} & \frac{\delta z}{\theta} \end{vmatrix} = \begin{vmatrix} sin(\phi)cos(\theta) & \rho cos(\phi)cos(\theta) & -\rho sin(\phi)sin(\theta) \\ sin(\phi)sin(\theta) & \rho cos(\phi)sin(\theta) & \rho sin(\phi)cos(\theta) \\ cos(\phi) & -\rho sin(\phi) & 0 \end{vmatrix} = \rho^2 sin(\phi) \tag{8}$$

Therefore:

$$dxdydz = \rho^2 sin(\phi)d\rho d\phi d\theta \tag{9}$$

The independance that existed between x, y and z when building our model is not preserved when changing to spherical coordinates. This explains why the particles are not evenly spread on the sphere when changing the coordinates from cartesian to spherical.

We have built the spherical coordinates in the following way:

$$\begin{aligned} \rho &= \sqrt{x^2+y^2+z^2} \\ \phi &= arctan(\tfrac{y}{x}) \\ \theta &= arccos(\tfrac{z}{\sqrt{x^2+y^2+z^2}}) \end{aligned} \tag{10}$$

Points uniformly spread on a sphere:

To build n points uniformly placed on a sphere we must use the notion of golden angle. This sphere is called the fibonacci sphere.

We define the angles phi and theta to place the n points on the sphere as follows:

$$\begin{aligned} indices &= [0.5, 1.5, 2.5, ..., n-0.5] \\ \phi &= arccos(1 - 2 * \tfrac{indices}{n}) \\ \theta &= \pi * (1+\sqrt{5}) * indices \end{aligned} \tag{11}$$

If we then plot these points using by turning them to cartesian coordinates :

$$\begin{aligned} x &= cos(\theta) * sin(\phi) \\ y &= sin(\theta) * sin(\phi) \\ z &= cos(\phi) \end{aligned} \tag{12}$$

We obtain this figure :

We calculated the angular correlation function for a sample of 10 000 particules.

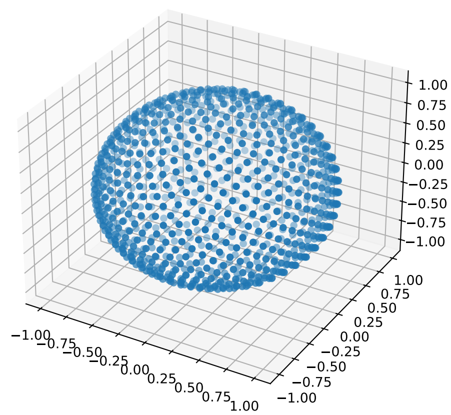


Figure 12: Fibonacci sphere : 1000 points uniformly distributed on a sphere.

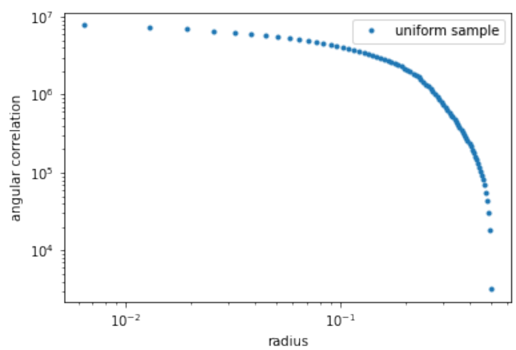


Figure 13: Angular correlation function for a sample of 10 000 particles following a uniform distribution.

3.2. *Gaussian distribution: one point and two points correlation functions*

This is not right because the transformation of gaussian particles to angular does not give a gaussian. We need to simulate gaussian par

4. ANGULAR CORRELATION FUNCTION

The angular correlation function is obtained using the estimator:

$$w(\theta) = \frac{DD(\theta)d\theta}{RR(\theta)d\theta} - 1 \tag{13}$$

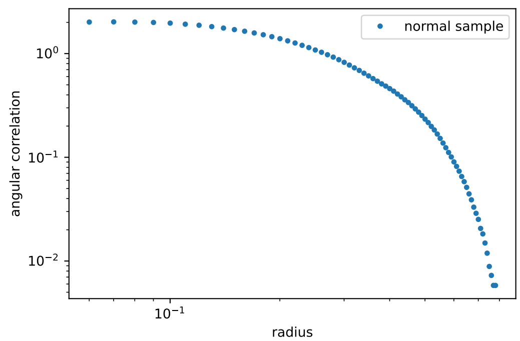


Figure 14: Angular correlation for a sample of 20 000 particles following a Gaussian distribution.

where $DD(\theta)$ is the number of pairs with angular separation $\theta + d\theta$ in the data and $RR(\theta)$ is the corresponding number in the random distribution.

Project your spheres onto the surface of the sphere which would mimick the two dimensional real observational catalogues. A paper to read is Landy and Szalay (Landy & Szalay 1993)

5. ANGULAR CORRELATION FUNCTION OF THE CATALOGUE OF QUASARS

The CatWise catalogue of quasars Secrest et al. (2021) contains around 1500000 quasars and it only gives the angular coordinates of each quasar on the sky. So we need do a angular correlation function etc. The data can be obtained here: Code and data available at https://doi.org/10.5281/zenodo.4431089

6. CONCLUSION

We have through this work created functions to calculate the pair correlation between points given by cartesian coordinates in a box and the angular correlation function for points on a sphere. The efficiency of the code is not optimal but it makes it possible to get a first idea on the nature of the data.

We applied these to the catalogue of quasars.

If we wanted to go further we could do the same work on the three dimensional data containing also the redshift associated to each entity. This model would be closer to reality but more difficult and longer to calculate.

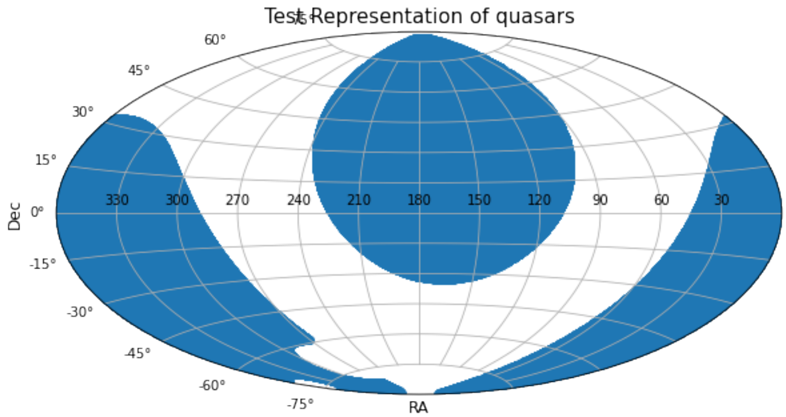


Figure 15: First visualisation of 1 350 000 quasars catalogue with coordinates Right Ascension and Declination (not galactic coordinates)

7. INTERESTING WEB SITES/READING REVIEW MATERIAL

Link to github : https://github.com/AstridBenamou/Angular-correlation

http://www.physics.emory.edu/faculty/weeks//idl/gofr.html

Reading material:

(1) Van den Bosch in this overleaf

REFERENCES

Landy, S. D., & Szalay, A. S. 1993, ApJ, 412, 64, doi: 10.1086/172900

Secrest, N. J., von Hausegger, S., Rameez, M., et al. 2021, ApJL, 908, L51, doi: 10.3847/2041-8213/abdd40