Cosmic Structure Formation

Computer Assignment: Point Processes and Correlations

January 9, 2017

1 Introduction

This assignment is meant to introduce you to the clustering properties and analysis of a range of spatial (3-D) point processes and distributions.

The assignment involves two parts. The first is the generation of a point set in a cubic volume. You will have to generate a few different point sets. Subsequently, in the second part of the assignment you will have to compute and analyze the two-point correlation function of the generated point distributions.

2 the Two-Point Correlation Function

The discrete equivalent of the autocorrelation function $\xi(r)$ is the **two-point correlation function** $\xi_{12}(r)$. If a given point distribution represents a fair sampling of the underlying continuous distribution, the two-point correlation function ξ_{12} should be equal to the autocorrelation function $\xi(r)$.

In cosmology the two-point correlation function $\xi_{12}(r)$ of a homogeneous point process is follows on the basis of the excess probability of finding points at a distance r. For a homogenous Poisson process one knows that if we take two volumes dV_1 and dV_2 at a distance r, the probability dP_{12} (or, rather,

number) of points in the two volumes is given by

$$dP_{12} \;=\; \bar{n}^2\, dV_1 dV_2\,. \tag{1}$$

For an inhomogeneous point process, i.e. in the case of clustering (due to the existence of underlying density perturbations), there will be an excess with respect to the Poisson distribution. This is encapsulated in the function $\xi_{12}(r)$,

$$\boxed{dP_{12} \;=\; \bar{n}^2\, \left\{1 + \xi_{12}(r)\right\}\, dV_1 dV_2} \tag{2}$$

In other words, the correlation function measures the excess probability. If there is clustering at a distance r, $\xi(r) > 0$. If points are anticorrelated at that distance, ie. tend to avoid each other, then $\xi(r) < 0$. And if there is no clustering at all but a homogeneous distribution we have $\xi(r) = 0$. Note that from now on we simply assume that $\xi_{12}(r) = \xi(r)$. Notice that we assume that because of the **isotropy** of the density fluctuations the two-point correlation function should also be isotropic and only a function of distance r. The significance of the two-point correlation function $\xi(r)$ has

formed the main tool in the study of the large scale galaxy distribution. It has formed the main statistical measure for clustering in the Universe. Every catalogue of galaxy positions, on the sky or in redshift space, has been analyzed to determine the two-point correlation function. The same holds true for catalogs of clusters of galaxies, of active galaxies, etc. There are a variety of reasons for its prominence:

- Clustering of galaxies, clusters of galaxies, radio galaxies, etc. is clearly an important aspect of the cosmic large scale matter distribution. The two-point correlation function is the first order measure for characterizing deviations from a uniform distribution: it forms the first order description of clustering.

- The autocorrelation function is the Fourier transform of the Power Spectrum $P(k)$, and in particular in the linear regime it contains crucial information on the cosmological scenario prevailing in our Universe. Hamilton et al. (1991) even managed to find a relation between the measured nonlinear $\xi(r)$ and the linear power spectrum.

- For highly nonlinear clustering we often find that the two-point correlation function is a power-law function of distance r,

$$\boxed{\xi(r) \;=\; \left(\frac{r}{r_0}\right)^{-\gamma}} \tag{3}$$

  The socalled **correlation length** r_0 (the name is a misnomer and often confusing for physicists, who have another definition) is a measure for the amplitude of the clustering process. It is the value of the distance at which $\xi(r) = 1$, and thus the distance at which the clustering strength becomes comparable to the probability of the homogeneous point process. It therefore provides a good measure for the **scale of nonlinearities**: above the correlation scale the point distribution rapidly enters the linear clustering regime.

- The corresponding **power-law slope** γ appears to have a rather universal value of $\gamma \approx 1.8$. In the nonlinear clustering regime γ is closely coupled to the slope n of the power spectrum $P(k)$,

$$n(k) \;\equiv\; \frac{d\log P(k)}{d\log k}\,, \tag{4}$$

  and thus contains a wealth of information on the underlying structure formation process.

- The most reliable estimates of the two-point correlation function concern the analysis of (two-dimensional) sky distributions of galaxies. The best galaxy sky catalogues contain millions of galaxies. Statistically this guarantees estimates with small errors. The resulting angular two-point correlation function $\omega(\theta)$ is basically a weighted projection of the spatial two-point correlation function $\xi(r)$ (expressed through the socalled Limber equation). On small scales, the power-law behaviour of the latter thus translates into a power-law angular two-point correlation function,

$$\omega(\theta) \;=\; \left(\frac{\theta}{\theta_0}\right)^{1-\gamma} \tag{5}$$

where γ is the power-law slope of the spatial two-point correlation function. Interesting is the behaviour of the angular correlation scale θ_0. It is very sensitive to the selection of galaxies in the catalogue: it scales with the depth of the sample. The large the apparent magnitude limit m_{lim}, i.e. the deeper we look into the Universe, the smaller θ_0 becomes. This of course is due to the projection of ever more shells on top of each other, as well as in a shift of the angular scale corresponding to a particular physical scale. There is a very precise relation between this angular correlation scale and the depth of the survey on the condition that **we live in a Universe which on the largest scales is homogeneous**. This indeed appears to be true, one of the most convincing arguments for the **Homogeneity of the Universe**, one of the basic tenets of the **Cosmological Principle**. Fairness demands to say that this finding has been challenged by a few groups, although none of them came up with convincing evidence for the contrary.

- The two-point correlation function plays an important role in dynamical analysis of structure formation: the measured cosmic flows can be related to the matter distribution through the two-point correlation function ("cosmic virial theorem", although for this we also need the *three-point function*"). From this we may infer cosmological parameters. Most noteworthy in this is the determination of the two-point correlation function in redshift space: the anisotropic distortions induced by the influence of cosmic flows on the measured redshifts can be directly translated into an estimate of Ω_m.

3 Point Processes and Distributions

A point process is a form of stochastic or random process. It may be thought of as a set of random points in a space, with a certain probability defined over the same space. Formally, it should be called a point field, but let us just use both names as stochastic variation on the theme. In general, we can broadly distinguish two kinds of point process, homogeneous and inhomogeneous process. If the intensity is not a function of location $\mathbf{x}$, then we speak of a homogeneous poisson process.

We restrict ourselves to point processes in three-dimensional space $\mathcal{R}^\ni$, and to the bounded region V in which a point field is located. In our case this will be the unit cube. Note that many other volumes may be imagined, for example the observationally defined pie slices.

For astronomy point processes are very relevant, many observables may be modelled by them. For example the spatial distribution of stars and galaxies can be thought of as a point process. In this respect we should note that in general the mean of the distribution is defined as the average value at a certain point in space over many realizations. Only if one assumes Ergodicity, the spatial mean is equal to the average of the probability distribution. Ergodicity is of utmost importance to cosmology as our Universe is the only one sample we have.
In this computer assignment, We will be looking at four different spatial point processes.

3.0.1 Random/Poisson distribution in Cubic Volume

One of the simplest and fundamental point processes is the spatial Poisson point process. The points are stochastically independent and the probability of the number of points N(A) in a region A, is given by the Poisson distribution:

$$P(N(A)) \;=\; \frac{(\lambda V(A))^k}{k!}\, e^{-\lambda V(A)} \tag{6}$$

Here λ is the intensity of the point distribution, which is the mean of the distribution.

Generating a Poisson field is straightforward. For a realization in a volume $V(A)$, determine with a random poisson deviate, the number of points lying in $V(A)$. Distribute these points randomly over the volume. In practice, it

is often assumed that the number of points in a volume $V(A)$ is so large that $N(A)$ can be taken as constant defining the total number of points in the sample volume, N. Strictly speaking, this is not correct, but for our purpose this you are advised to do.

For the generation of the random points, use a uniform random number generator $U[0,1]$. A computer random number generator only returns a uniformly distributed random number $Z \in [0,1]$. For a d-dimensional space, for a Poisson process in a Cartesian coordinate system you generate one random number for each coordinate.

In summary, when you wish to generate a random point $\vec{X}_i = (x_i, y_i, z_i)$ in a cubic volume with size L, with all points contained in its volume,

$$V = [0, L] \times [0, L] \times [0, L]\,, \tag{7}$$

you can obtain this via calling three times the random number generator to obtain three uniformly distributed numbers Z_1, Z_2 and Z_3 ($Z \in [0,1]$),

$$\begin{aligned} x_1 &= Z_1 L\,, \\ y_i &= Z_2 L\,, \\ , z_i &= Z_3 L\,. \end{aligned} \tag{8}$$

Note that if you generate random points in other coordinate systems, such as spherical coordinates, you have to take care that the volume elements are properly accounted for.

3.0.2 Random/Poisson distribution in Sphere

For the generation of a uniformly distributed point set in a sphere of radius R there are a few options.

By far the most efficient method is to generate the spherical coordinates (r, θ, ϕ) of a point. For spherical coordinates you have to take into account the volume element defined by a range $(dr, d\theta, d\phi)$ in the spherical coordinate system,

$$dV \;=\; r^2 \cos\theta dr d\theta d\phi \tag{9}$$

To generate a randomly distributed point in the spherical volume, generate three random numbers $(Z_1, Z_2, Z_3 3) \in [0,1]$ from a uniform distribution

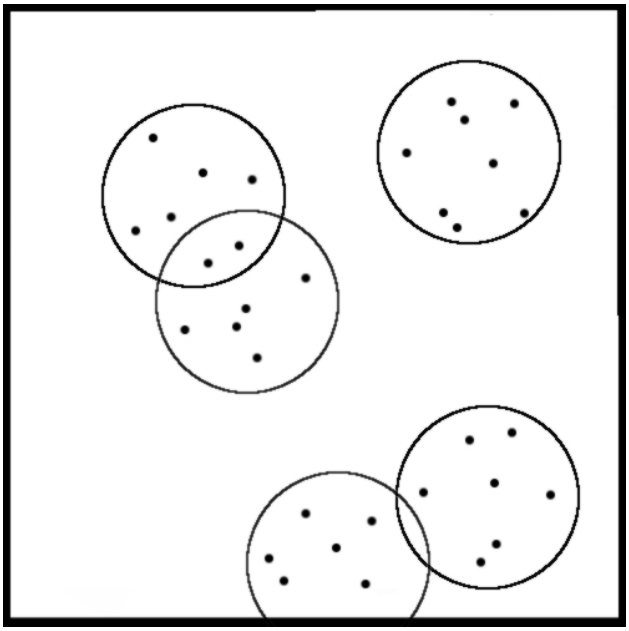


Figure 1: Illustration of the generation of a Matern process.

$U[0,1]$. Subsequently, you can obtain the spherical coordinates $(r_i, \theta_i \phi_i)$ of the random point i as follows:

$$\begin{aligned} r_i &= R\, Z_1^{1/3}\,, \\ \theta_i &= \arccos\left(1 - 2Z_2\right), \qquad (10) \\ \phi_i &= 2\pi\, Z_3\,. \qquad (11) \end{aligned}$$

By repeating the procedure above N times, you obtain a nice distribution of N uniformly distributed points within a sphere of radius R, with Cartesian coordinates $\vec{X}_i = (x_i, y_i, z_i)$,

$$\begin{aligned} x_i &= r_i \sin\theta \cos\phi\,, \\ y_i &= r_i \sin\theta \sin\phi\,, \qquad (12) \\ z_i &= r_i \cos\theta\,. \end{aligned}$$

3.0.3 Matern process

In a Matern process, spheres of fixed size R are generated. Each sphere is subsampled with a Poisson point distribution, with mean μ. For a realization

see Figure 1. The generating process consists of three steps,

1. Decide on number N of randomly located (spherical) clusters.

2. For each cluster draw a random position $\vec{C}_i$, the center of a (spherical) cluster i.

3. Populate each sphere i uniformly (!) with a Poisson random number N_i of points. This number follows from the Poisson distribution with mean μ and volume $V_i(R)$ of the sphere (see eqn. 6),

$$V_i(R) \, = \, \frac{4\pi}{3} \, R^3 \, . \tag{13}$$

3.0.4 Soneira-Peebles model

The Soneira-Peebles model is a fractal-like pont distribution involving hierarchically embedded levels of ever larger point density, see Figure 2a. It was introduced by Soneira and Peebles to model the galaxy distribution obeying various clustering measures. A realization is generated as follows:

1. The starting point is a level-0 sphere of radius R.

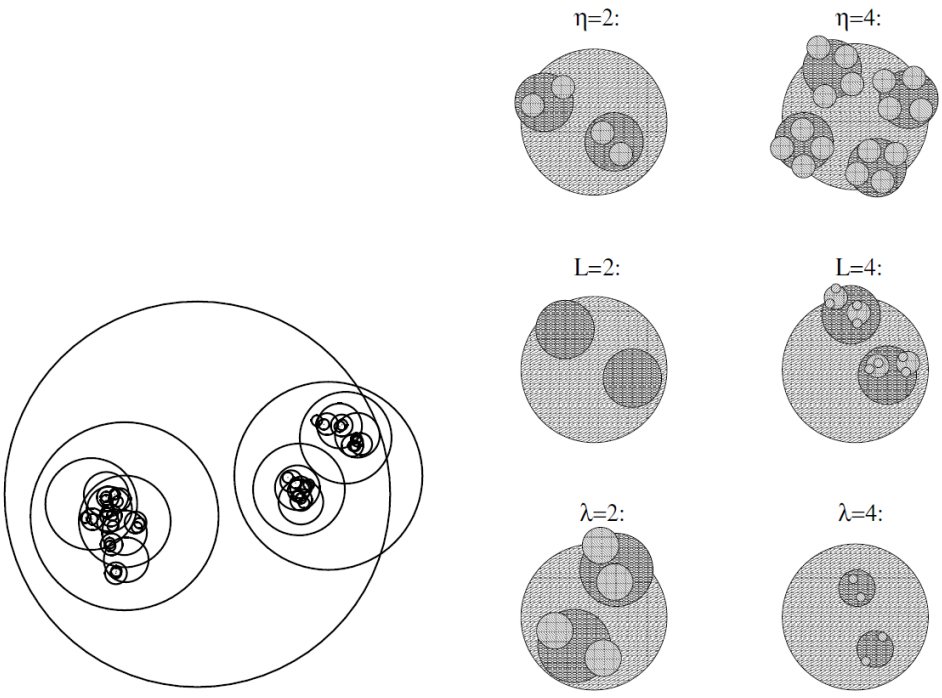
2. In this sphere η level-1 spheres are placed with radius R/λ and $\lambda > 1$. The new spheres are placed at a random position inside the level-0 circle, such that their centers fall inside the original level-0 sphere.

3. Within each of these η level-1 spheres, one places η level-2 spheres of radius R/λ^2.

4. This process is repeated until one ends up with in total η^L level-L spheres of radius R/λ^L. At the center of each of these level-L spheres a point is placed.

One therefore ends up with in total η^L points, which in the Soneira-Peebles model represent galaxies. This procedure is illustrated in the top panel of Figure 2.

The Soneira-Peebles model is controlled through three parameters, μ, L and λ. The effect of varying these parameters on the resulting point distribution is illustrated in the 2nd to 4th row of Figure 2.



(a) The Soneira-Peebles model. Inside a level-0 sphere η level-1 circles are placed with a radius which is smaller by a fixed factor. This process is repeated until one ends up with η^L level-L circles. At the center of these level-L circles η^L points are placed, which form the resulting Soneira-Peebles point distribution.

(b) The physical meaning of the three defining parameters η, L and λ of the Soneira-Peebles model. The upper row shows the effect of varying η, the number of circles which is placed in each circle. The central row shows the effect of varying L, the total number of levels. The bottom row shows the effect of varying λ, the ratio of the radius of each circle with the radius of subsequent circles of one level higher.

Figure 2: Definition of the parameters of the Soneira-Peebles process.

For a given number of points, η determines the dynamic range of the resulting point distribution. For a small value of η, many levels are needed to reach a fixed number of points, while a large value of η results in a smaller number of levels. A small value of η also results in a smaller filling fraction of space with spheres than a high value of η (2nd row in Figure 2). L denotes the total number of levels and therefore determines the range of densities and scales in the resulting point distribution. For a fixed value of η, L also determines the total number of points (third row in Figure 2).

Finally, for given values of η and L, λ determines the range of spatial scales. A value of λ close to 1 means that subsequent spheres of higher levels are of comparable size. Values of λ much larger than one mean that each

subsequent level consists of spheres which are significantly smaller than the spheres in the preceding level (bottom row in Figure 2).

An important property of the Soneira-Peebles model is that it is one of the few analytic self-similar models of the galaxy distribution for which the two-point correlation function can be analytically evaluated.

4 Assignment/computer task, part1: the Point Samples

The point samples to be generated should be contained in a cubic volume of size L, with volume $V = L \times L \times L$.

In all cases it concerns a cube with periodic boundary conditions. This means that the cube is surrounded by copies of itself. In other words, each point in the cube is also found in each of the surrounding 26 cubes (the copies are translated by the appropriate amounts of L, e.g. a point (x, y, z) has a periodic copy $(x + L, y, z)$ in the cube on the right and a copy $(x + L, y + L, z + L)$ in the cube on the upper righthand side.

Armed with this knowledge, this assignment concerns the following:

1. Write a program that generates a uniform point distribution with $N = 10000$ points in a cube of size L. Plot the positions of the points on the xy plane (ie. plot the x-coordinate vs. y-coordinate of each point.

2. Write a program that generates random points in a sphere of radius R. For a sphere of unit radius and 10000 point, plot the positions of the points in the xy plane and the yz plane. Infer the radial distribution, ie. the distribution $f(r)$ of the radial distance of each point to the center of the sphere.

3. Write a program that produces a realization of the Matern point distribution, and plot the generated point distribution in the xy plane for the parameters:

   a. $\lambda = 1000$, $r = 0.05$ and $N = 12$

   b. $\lambda = 1000$, $r = 0.05$ and $N = 100$

4a. MSc students: write a program that produces a realization of the Soneira-Peebles point distribution, and plot the generated point distribution in the xy plane for the parameters:

   a. $\eta = 6$, $\lambda = 3$, $L = 6$

   b. $\eta = 3$, $\lambda = 1.7$, $L = 10$

   c. $\eta = 4$, $\lambda = 1.9$, $L = 8$

4b. BSc students: write a program that produces a realization of the Soneira-Peebles point distribution, and plot the generated point distribution in the xy plane for the parameters:

a. $\eta = 6$, $\lambda = 3$, $L = 6$

5 Measuring the Two-Point Correlation Function

The issue is of course how to measure the clustering properties of the various point processes by means of the correlation function.

Again, we are beset by the problem that there is only one realization of our Universe known. Our own cosmos. Luckily, also here we are saved by the *ergodic theorem*. We may measure the function by averaging over many different positions. Thus, we will follow this approach. In essence, it becomes a large counting exercise. We are going to count the number of points within spherical shells around a given point. By adding them all up and averaging them in a proper way we get an estimate of the probability that on average at a distance r we have a certain amount of points and thus it's excess or deficit with respect to a homogeneous Poisson process. From this we may infer $\xi(r)$. Easier said then done ...

5.1 Correlation Function Estimators

For measuring $\xi(r)$ on the basis of point counts we need to take into account that we cannot always fit in complete spheres of radius r at every position within a survey volume. In other words, one needs a way of dealing with *edge corrections*. The common practice is to deal with this by means of an equivalent Poisson point catalog in exactly the same volume (and with the same selection criteria concerning depth of the survey).

Assume we have therefore two point sets. One is the sample one, designated by the letter "D" (data). It contains N_D points. In addition there is the Poisson point set "R", with N_R points. Position yourself on a number (as large as practically feasible) of the data points and count the number of data points you find in a spherical shell of with radius $[r, r + \Delta r]$. The total sum of points counted is designated as $DD(r)$. One may also count, for the same number of points, the number of Poisson points in the same shells, $DR(r)$.

- The first estimator is that defined by Davis and Peebles (1983), sometimes called the *standard estimator*,

$$\hat{\xi}_{DP}(r) = \frac{N_R}{N_D} \frac{DD(r)}{DR(r)} - 1 \tag{14}$$

- Hamilton (1993) found systematic biases in this estimator, surpassing the regular uncertainties (due to finite sampling) in $\hat{\xi}_{DP}(r)$. He therefore proposed the socalled *Hamilton estimator*:

$$\hat{\xi}_{HAM}(r) \;=\; \frac{DD(r)\cdot RR(r)}{[DR(r)]^2} \;-\; 1 \tag{15}$$

  in which $RR(r)$ is the number of pairs in the random catalog with separation in the interval $[r, r+\Delta r]$.

- Almost simultaneously another improved estimator was defined by Landy & Szalay (1993). It has similar properties as the Hamilton estimator,

$$\hat{\xi}_{LS}(r) \;=\; 1 \;+\; \left(\frac{N_R}{N_D}\right)^2 \frac{DD(r)}{RR(r)} \;-\; 2\frac{N_R}{N_D}\,\frac{DR(r)}{RR(r)}\,. \tag{16}$$

6 Assignment/computer task, part 2: point clustering & correlation functions

In this experiment you will be invited to determine the two-point correlation function of each of the point sets that you have generated in section 4.

In this second part of the assignment, you will have to measure the two-point correlation function of these point sets, produce linear-linear plots and log-log plots, and determine the correlation length r_0 and clusterig slope γ of the 2-pt correlaation function.

We assume, for simplicity, that there these are completely sampled point sets without any sampling involved (such as involving a radial selection function).

- Write a program to analyze your point set and make the corresponding graphs of the function. You will have to do this both in logarithmic bins, in particular at small scales, to infer the predicted power-law shape (log-log plot). In addition, for the same set you will have to determine $\xi(r)$ out to somewhat larger scales in linear bins (lin-lin plot).

- As at small distances the two-point correlation function behaves like a power-law of the distance r between points,

$$\xi(r) \;=\; \left(\frac{r}{r_0}\right)^{-\gamma} \;, \tag{17}$$

  you will also have to fit a power-law to the function you have produced. From this fit you should derive the **power-law slope** γ and the **correlation length** r_0.

- MSc students: Repeat this for three different estimators $\hat{\xi}$ of the two-point correlation function, the "standard" *Davis-Peebles estimator*, the *Hamilton estimator* and the *Landy-Szalay estimator*. See the following subsection for their specification.

- BSc students: Use the the "standard" *Davis-Peebles estimator* $\hat{\xi}$ for the calculation of the two-point correlation functions of the generated point sets.