1 Introduction

It is well known that the distribution of galaxies on the sky is far from uniform. Qualitatively, the non-uniformity has been described in terms such as clusters, clusters of clusters, filaments, voids, etc. Large-scale surveys, where the positional information (coordinates on the sky) have been complemented with depth information from redshifts, have mapped the 3D structure of this distribution in exquisite detail. In this assignment we shall however only deal with the projected 2D (or surface) distribution of galaxies on the sky.

Whether we want to investigate the 3D or 2D distribution of galaxies (or of any other kind of object), a fundamental problem is how to quantify the non-uniformity. Whereas a completely uniform (random) distribution is characterized by a single number (the mean density of objects, i.e., per unit surface or unit volume), there are an infinite number of possible ways in which a non-uniform distribution could be quantified (modelled, or parameterized). The two-point angular correlation function $w(\theta)$ is one, very useful, way to do this. It is described in Chapter 10.4 of the textbook [3].

2 Project

In this assignment, six different datasets will be used. Each dataset is a list of 9404 positions ($x$ and $y$ coordinates) of points in a square field. The coordinates have been scaled such that $0 < x < 1000$ and $0 < y < 1000$. The datasets can be retrieved from the ASTM21 web page as text files P1data01.txt, P2data02.txt, etc.

One of the datasets contains the measured positions of galaxies in the Hubble Ultra Deep Field (HUDF) [1], as observed in the $i$ (F775W) band, and therefore exhibits a certain degree of clustering on some scales. The other five datasets were randomly generated with a uniform distribution, and therefore by definition has no clustering.

The task is to estimate the two-point correlation function $w(\theta)$ for each dataset and hence decide which is the galaxy dataset. The angle $\theta$ (expressed in the same unit as the $x$ and $y$ coordinates) ranges from 0 to $\approx 1400$ units.

It is recommended to estimate $w(\theta)$ only for $\theta \leq \theta_{\text{max}} \approx 1000$, since for larger separations there are too few pairs to get reliable statistics.

Two different estimators should be used and compared: the ‘natural’ estimator ($w_1$) and the Landy–Szalay estimator ($w_3$). Once the galaxy field has been identified, describe how it deviates from a uniform distribution on different scales.
3 Theory

The basic formulae are given in the textbook [3], Eqs. (10.7)–(10.11). Details (including the derivation of \( w_3 \)) can be found in the paper by Landy & Szalay [2]. For convenience, the formulae for \( w_1 \) and \( w_3 \) are given hereafter.

The dataset under investigation (i.e., one of P1data*.txt) is denoted \( D \) and contains \( n = 9404 \) points. The two-point correlation function is obtained by comparing the distribution of angular separations in \( D \) with the corresponding distribution of separations in a random set, called \( R \), and which contains \( r \) points. In general \( r \) and \( n \) may be different, but it is often convenient to choose \( r = n \).\(^1\)

The angular separation of two points \( i, j \) is \( \theta_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \). The total number of pairs in \( D \) and \( R \) is \( n(n-1)/2 \) and \( r(r-1)/2 \), respectively. To characterise the clustering at the angular scale \( \theta \), count the number of pairs in \( D \) and \( R \) that have separations in the interval \( \theta \pm \delta \theta/2 \). Let \( DD \) and \( RR \) be the number of such pairs in \( D \) and \( R \), respectively. The ‘natural’ estimator of \( w(\theta) \) is

\[
 w_1 = \frac{DD}{n(n-1)/2} - \frac{RR}{r(r-1)/2} = \frac{r(r-1)}{n(n-1)} DD - 1. \tag{1}
\]

To reduce the effect of statistical fluctuations in \( RR \) it is advisable to generate several (say, 10) random fields, and use the average number \( \langle RR \rangle \) in the above formula.

As shown by Landy & Szalay [2], \( w_1 \) is not optimal because its variance decreases relatively slowly with increasing numbers of data points (i.e., even when there is no clustering, it tends to give values that are far from 0 unless there are very many data points). This makes it harder to see if there is real clustering in the dataset. A much better estimate is obtained by using also the cross-correlation statistic \( DR \), that is the number of pairs in the separation interval \( \theta \pm \delta \theta/2 \), with one point taken from \( D \) and the other from \( R \).

There are \( nr \) such pairs. The Landy–Szalay estimator is then computed as\(^2\)

\[
 w_3 = \frac{DD}{n(n-1)/2} - \frac{2DR}{nr} + \frac{RR}{r(r-1)/2} = \frac{r(r-1)}{n(n-1)} DD - \frac{r-1}{n} DR + 1. \tag{2}
\]

It is advisable to replace \( RR \) and \( DR \) by the mean counts from several realizations of \( R \).

In the actual calculations, \( DD \), \( RR \), and \( DR \) should be arrays of length \( m \) (\( \approx 100 \)), counting the number of pairs in \( m \) bins of equal width. For example, \( DD(1) \) is the number of pairs with \( 0 < \theta \leq \delta \theta \), \( DD(2) \) the number of pairs with \( \delta \theta < \theta \leq 2\delta \theta \), and so on, up to \( DD(n) \), which is the number of pairs with \( (m-1)\delta \theta < \theta \leq m\delta \theta \equiv \theta_{\text{max}} \). Here, \( \delta \theta = \theta_{\text{max}} / m \) is the bin width.

As this project is computationally rather intensive, you should be careful so that you do not repeat calculations unnecessarily. This is often a question of doing nested loops in the right order. For example, rather than looping through the bins and checking which pairs fall in that bin, it is better to loop through all the pairs only once, calculate the index of the bin to which the pair belongs, and add 1 to the count in that bin.

\(^1\)To be sure that \( R \) is truly uniform random, you have to generate it yourself; do not use one of the other datasets (which may or may not be random)! Use \( r = n = 9404 \).

\(^2\)Here, and in [3], \( DD \), \( RR \), and \( DR \) denote the number of pairs in the bin, while in [2] they denote the normalized counts, i.e., divided by the total number of pairs.
4 Reporting

Students are encouraged to discuss and work together on the project. However, each student must produce a written report with his/her own solution, including a description of the method, results in the form of diagrams etc, and conclusions. The report should be sent as a pdf file to paul@astro.lu.se, together with the complete code (MATLAB scripts/functions) written by the student and used in the project.

References

