Bull. Astr. Soc. India (2005) **33**, 1–9

Fractal analysis of galaxy surveys

T. R. Seshadri*
Department of Physics and Astrophysics,
University of Delhi, Delhi 110 007, India

Received 21 March 2005; accepted 29 March 2005

Abstract. Fractal analysis is a powerful tool to study the nature of galaxy distribution. The use of multifractal analysis of the galaxy distribution for the investigation of the transition to homogeneity in the Universe is reviewed. This analysis shows that the Universe is homogeneous over scales larger than about $80h^{-1}$ to $100h^{-1}$ Mpc.

Keywords : cosmology: miscellaneous – cosmology: theory – methods: statistical – large-scale structure of the Universe

1. Introduction

The framework of Cosmology is built on the hypothesis of homogeneity and isotropy of the Universe at least over very large scales. From a philosophical point of view this claim is appealing because it implies that there is no special point in the Universe. So attractive is this hypothesis that it has become one of the central pillars for modern cosmology. However, observational verification of this hypothesis is essential before one can consider seriously the various aspects of cosmology today that are based on this fundamental hypothesis.

The distribution of galaxies, as indicated by galaxy surveys, suggests that the Universe is far from being homogeneous and isotropic. There are regions where one finds aggregation of galaxies. On the other hand, one also finds regions, called voids, that are conspicuous by the sparcity of galaxies. As we probe galaxies with larger redshifts, we are probing their distributions over larger scales in the Universe. There have been several surveys of galaxies conducted over the years. Initially the surveys involved only angular distributions. We now have surveys which have information about the distances of the galaxies from us as well. Thus one can infer the three dimensional position of a large number of galaxies.

*e-mail:trs@physics.du.ac.in

One of the initial surveys of this kind was the *CfA* survey. The *CfA-II* survey probed the Universe to a distance of about $150h^{-1}$Mpc. For $h = 0.71$ (as indicated by the results of *WMAP*), this comes out to be about 200Mpc. Over the years one has been able to probe farther distances in the Universe. The Las Campanas Redshift Survey (Shectman et al. 1996) probes distances up to about 850Mpc for $h = 0.71$. Other surveys like the Sloan Digital Sky Survey, probe the Universe to even farther distances.

Another technique which probes the distribution of matter, *albeit* more indirectly, is the analysis of the Cosmic Microwave Background Radiation (*CMBR*). Observations indicate that the *CMBR* has a very high degree of isotropy. The photons of *CMBR* which we receive today contain information about the state of the Universe at a redshift of $z \sim 1000$. The epoch of the transition from the plasma era to the neutral phase is denoted by t_{rec}. After this epoch, the photons by-and-large decouple from matter and travel almost unscattered till they reach us today t_0. After these photons got scattered for the last time, they would have travelled a certain distance before reaching us. This distance would be the same in all directions. We can then define a hypothetical spherical surface of last scatter with us in the centre of that (hypothetical) sphere. The isotropy of the *CMBR* implies that all the points on the spherical surface are similar. We next assume that there is nothing special about our position in the Universe. Any observer (located at any other point) would see the same isotropy (statistically). Hence continuing this chain of reason, all spherical surfaces would look the same whereever they are placed. This would be possible only if the Universe is homogeneous.

In view of these two limits, namely those of homogeneity over very large scales and inhomogeneity over relatively smaller scales (as indicated by the galaxy distribution in our neighbourhood), it is natural to investigate the length scale over which this transition from inhomogeneity to homogeneity takes place. The overall evolution of the Universe assumes an underlying homogeneity and isotropy. This would make sense only if applied to length scales which are bigger than those corresponding to this transition and a measure of this transition scale is very crucial for Cosmology.

Correlation function analysis has been a common method for this investigation. Fractal analysis is another method of investigating the nature of this transition and in this review, we focus on this aspect.

2. Galaxy surveys and the motivation for fractal analysis

Correlation function measures the conditional probability of finding a galaxy at a position $\mathbf{x} + \mathbf{r}$, given that there is another galaxy at a position $\mathbf{x}$. More precisely, this is called the 2-point correlation function. In practice, the information we have from the surveys is not that of probability but of positions of galaxies. In order to implement this definition, we assume that the number density of the galaxies is proportional to the probability density of finding a galaxy at a point. If the galaxies do not show any tendency of clustering, the probability density of finding a galaxy at a point is independent of finding another galaxy at any other point. In terms of the number

density of galaxies, this would mean,

$$\langle n(\mathbf{x})n(\mathbf{x}+\mathbf{r})\rangle_{\mathbf{x}} = n_0^2 \tag{1}$$

where, the subscript $\mathbf{x}$ implies that the term in the angular brackets has been averaged over all values of position ($\mathbf{x}$) and n_0 is the average number density of the galaxies, i.e., $n_0 = \langle n(\mathbf{x})\rangle_{\mathbf{x}}$. In practice, however, the averaging cannot be done over all the positions. We can only do it over the region spanned by the survey. This is a limitation of the correlation function approach.

Clustering of points, (that of galaxies in our context) is characterized by the fact that the conditional probability is greater than the product of the probability of finding a galaxy at these two points. In terms of the number density of galaxies, it translates to,

$$\langle n(\mathbf{x})n(\mathbf{x}+\mathbf{r})\rangle_{\mathbf{x}} > n_0^2 \tag{2}$$

which can alternatively be expressed as,

$$\langle n(\mathbf{x})n(\mathbf{x}+\mathbf{r})\rangle_{\mathbf{x}} = n_0^2(1+\xi(\mathbf{r})) \tag{3}$$

The quantity ξ is defined as the 2-point correlation function and can be written as,

$$\xi(\mathbf{r}) = \frac{\langle n(\mathbf{x})n(\mathbf{x}+\mathbf{r})\rangle_{\mathbf{x}}}{n_0^2} - 1 \tag{4}$$

It is worth emphasizing that the process of averaging, which should in principle have been done over all regions in the Universe, can in practice be done only within the sampled region. This would be equal to the average number density of galaxies in the Universe only if the transition to homogeneity occurs within the sample volume, i.e., the length scale of the transition is smaller than the linear scale of the sample. This certainly is not guaranteed. This feature is a drawback of the correlation analysis.

The two point correlation function $\xi(r)$ is very well determined on small scales (Peebles 1993 and references therein). It scales as a power law in the distance between the two points.

$$\xi(r) = \left(\frac{r}{r_0}\right)^{-\gamma} \quad \text{with} \quad \gamma = 1.77 \pm 0.04 \quad \text{and} \quad r_0 = 5.4 \pm 1h^{-1}\text{Mpc} \tag{5}$$

This power-law form suggests a scale invariant behaviour on scales less than r_0.

Another statistical measure of the nature of clustering is the fractal dimension. An advantage of this is that we do not have to assume that the homogeneity scale is less than the linear scale of the survey. Further, as we will see below, the multi-fractal analysis contains much more information than the simple two-point correlation function.

3. Fractional dimensions and fractals

Fractals are an interesting way to characterize a distribution of points (Feder 1989; Falconer 1990; Borgani 1995). While studying the nature of clustering of galaxies, we consider these galaxies as points distributed in space. Fractals are characterized by a parameter called fractal dimension. In general, there is not merely a single parameter but a spectrum of parameters that characterize this distribution. As we shall see, the distribution of galaxies, up to a certain distance is characterized by such a spectrum.

3.1 Box-counting dimension

Consider a box and a set of points distributed in it. We now divide each side of the box into half. The number of smaller boxes will depend on the Euclidean dimension of the space in which the box is embedded. For a d-dimensional space, the number of smaller boxes will be 2^d. We are now interested in the number of boxes that contain at least one point from the sample. In other words, we need to count the number of non empty boxes. All the 2^d boxes need not necessarily have points in them. We then further divide each of these boxes into halves giving rise to 2^{2d} number of boxes and we again count the number of non-empty boxes. This procedure is repeated and at every stage we count the number of non-empty boxes, $N(r)$, where r is the linear size of the boxes. If $N(r)$ scales as a power law in r, in the limit $r \longrightarrow 0$, we call the exponent as the box-counting dimension D_{box}

$$D_{box} = -\lim_{r \longrightarrow 0} \frac{d \log N(r)}{d \log r} \tag{6}$$

We now justify calling this parameter as 'dimension'. Consider a smooth curve in, say, 3 dimensions. We follow the procedure outlined above to compute this parameter. We enclose it in a $3D$ box, divide the box into smaller and smaller (and hence, more and more) boxes and count the number of non-empty boxes at every stage. In the limit of the box-size tending to zero, only the infinitesimally small boxes along the curve will contribute to this number. Hence, in this limit the number of non-empty boxes will scale as $N(r) \propto r^{-1}$, thus giving the box-counting dimension to be 1 which is the same as our usual understanding of dimension of a smooth curve.

In a similar way, we can consider a surface in $3d$. Following the above procedure of considering a box and dividing it into smaller boxes, we find that the fractal parameter is 2 which is equal to the dimension of the surface. These are examples where the dimension is an integer. We can, however, envisage a situation, in which, there is a distribution of points, for which, the parameter which we call dimension, turns out to be a fraction. We call such distributions as fractals. In other words, for such distributions, $N(r) \propto r^{-d}$, where d is not necessarily an integer.

There is, however a problem in using the above definition in practice. In a physically relevant situation, we only have a countable set of points. In such a case, when we take the limit $r \longrightarrow 0$ the size of the boxes become smaller than the inter-particle separation. In this limit the

number of non-empty boxes do not change as we reduce the box-size further. In such a case, the box-counting dimension turns out to be zero in the limit $r \longrightarrow 0$. Hence, in practice, using the definition in this strict sense outlined above is of little value. Operationally we look for a reasonable range of r in which $N(r)$ scales as a power-law in r and call the exponent as the box-counting dimension.

3.2 Correlation dimension

Correlation dimension is another way of characterizing fractal distributions. For a distribution with N points we label the points using an index j which runs from 1 to N. Of the N points a subset of M points indexed by i are selected.

For every point i, we count the total number of points which are within a distance r from the i^{th} point. We denote this number by $n_i(r)$, where,

$$n_i(r) = \sum_{j=1}^{N} \Theta(r - \mid \mathbf{x_i} - \mathbf{x_j} \mid) \tag{7}$$

where $\mathbf{x_i}$ is the position vector of the i^{th} point and $\Theta(x)$ is the Heaviside function. $\Theta(x) = 0$ for $x < 0$ and $\Theta(x) = 1$ for $x \geq 0$. Clearly $n_i's$ will be different for different $i's$. The probability $p_i(r)$ of finding a particle at a distance $\mathbf{r}$ about the centre i is obtained by dividing $n_i(r)$ by N. Since we are interested in the statistical features of the distribution, we further average $n_i(r)$ over the set of M centres. The resulting quantity, $C_2(r)$ is given by,

$$C_2(r) = \frac{1}{MN} \sum_{i=1}^{M} n_i(r)\,. \tag{8}$$

If the C_2 varies as a power-law in r

$$C_2(r) \propto r^{D_2} \tag{9}$$

the parameter, D_2, is called the correlation dimension. In a real situation as in the case of galaxy distribution, the value of D_2 may be different over different ranges of scale.

The quantity $C_2(r)$ is closely related to the correlation function that we discussed earlier. This quantity is related to the volume integral of the two point correlation function. For the regime, where the two point correlation function exhibits a power-law behaviour $\xi(r) = (r/r_o)^{-\gamma}$, we expect the correlation dimension to have a value $D_2 = 3 - \gamma$.

4. Multi-fractals

The box-counting dimension and the correlation dimension quantify different aspects of the scaling behaviour of a point distribution and they will have different values in a generic situation.

In fact one can define several kinds of dimensions by considering moments of the distribution. All these different types of dimensions are contained in the concept of generalized dimension. It provides a continuous spectrum of dimensions D_q for a range of the parameter q. One of the ways one can define generalized dimension is the Minkowski-Bouligand dimension D_q (Feder 1989; Falconer 1990). It is defined on similar lines as the correlation dimension. The difference, however, is that we use the $(q-1)$th moment of the galaxy distribution $n_i(r)$ (eq. 7) around any point. We first construct

$$C_q(r) = \frac{1}{NM}\sum_{i=1}^{M}[n_i(<r)]^{q-1}\,. \tag{10}$$

The generalized dimension is given by,

$$D_q = \frac{1}{q-1}\frac{d\ln C_q(r)}{d\ln r}\,. \tag{11}$$

The quantity $C_q(r)$ may exhibit different scaling behaviour over different scales and we can in principle have more than one spectrum of generalized dimensions over these different ranges.

As is clear from from equations (10) and (11) the generalized dimension D_q corresponds to the correlation dimension at $q = 2$.

Coinsider the case for $q = 1$. The exponent of n_i (which is $q-1$) in equation (10) is zero. The quantity n_i^0 is zero for $n_i = 0$ and is equal to unity for $n_i \neq 0$.

$$\begin{aligned} n_i^0 &= 0 \quad for \quad n_i = 0 \\ n_i^0 &= 1 \quad for \quad n_i \neq 0 \end{aligned} \tag{12}$$

Note that the implication of the above is that, if $n_i \neq 0$, the actual value of n_i is irrelevant in equation (10). In other words, the sum appearing in equation (10) counts simply the number of non-empty boxes, without any reference to the actual number in those boxes. This is precisely the way we have defined box-counting dimension. Hence, D_q for $q = 1$ corresponds to the box-counting dimension.

For a mono-fractal the generalized dimension is a constant i.e. D_q is independent of q. It reflects the fact that for a mono-fractal, the point distribution is characterized by a unique scaling behaviour. Further, if D_q is also equal to the Euclidean dimension, the distribution is homogeneous. For a multi-fractal on the other hand, the values of D_q will be different for different values of q. For $q > 0$, D_q probes the over-dense regions like clusters. The negative values of q, on the other hand, give more weightage to the under-dense regions like voids. Minkowski-Bouligand generalized dimension D_q is one of the possible definitions of a generalized dimension (Borgani 1995). Another way to characterize multi-fractals is the minimal spanning tree used by van der

Weygaert and Jones (van der Weygaert & Jones 1992; Martinez 1991; Martinez & Jones 1990). The calculation of the Minkowski-Bouligand generalized dimension is computationally simpler than most other methods.

5. Multi-fractal analysis of galaxy surveys

While dealing with real galaxy surveys, the situation is much more complicated than it is for point distribution. While we can theoretically define the concepts of dimension and calculate them for a given distribution, there are several non-trivialities which come up when these are applied to Galaxy surveys. As in the case of point set, the procedure is to choose a galaxy and draw circles of increasing radii and count the galaxies in a particular circle as a function of the radius. (This is the case when we consider slice surveys. For $3D$ surveys, this method needs to be generalized by considering spheres instead of circles.) As long as we are sufficiently inside the survey region, the effect of boundaries do not affect our analysis. Difficulties arise when we choose the centre near the boundary of the sample. As we increase the radius of the circle beyond a particular value, a part of the circle falls outside the sample region. Clearly, the nearer one is to the boundary, the smaller should be the radius one considers. Operationally one can decide the largest radius one is interested in, and identify the region in the distribution so that if the circle of that radius is centred on any galaxy in that region, the circle should remain wholly inside the sample. The problem is that the larger the scale we intend to analyze, the smaller the zone of inclusion and smaller the number of centres available for averaging. This leads to larger statistical variations.

Further, one needs to apply several corrections on the observed data before one can subject it to the fractal analysis. For example, the fainter galaxies far away tend to get missed out in the survey and needs correction factors in order to compensate for this. Further, in the case of slice surveys, the slices are collapsed to a two dimensional sheet. From the geometry of the slice, it is clear that this process tends to include excess of galaxies for large distances as compared to the nearby regions. Using different correction factors, one assigns an effective number of galaxies corresponding to every galaxy observed. While calculating C_q for the galaxy distribution, it is this effective number which is to be used. If w_j is the correction factor for the j^{th} galaxy, the effective number of galaxies in a circle (for slice surveys) within a radius $\mathbf{r}$ about the centre i is given by,

$$n_i(r) = \sum_{j=1}^{N} w_j \Theta(r - \mid \mathbf{x_i} - \mathbf{x_j} \mid) \tag{13}$$

Using this expression for n_i, one calculates the behaviour of $C_q(r)$ with r for different q's (Bharadwaj et al. 1999; Pan & Coles 2000; Baryshev & Bukhmastova 2004).

6. Generic results

It is observed that the behaviour of C_q is different at large and small scales. In the analysis of Las Campanas Redshift Survey, for example, this change occurs at about $80h^{-1}$Mpc. Figure 1

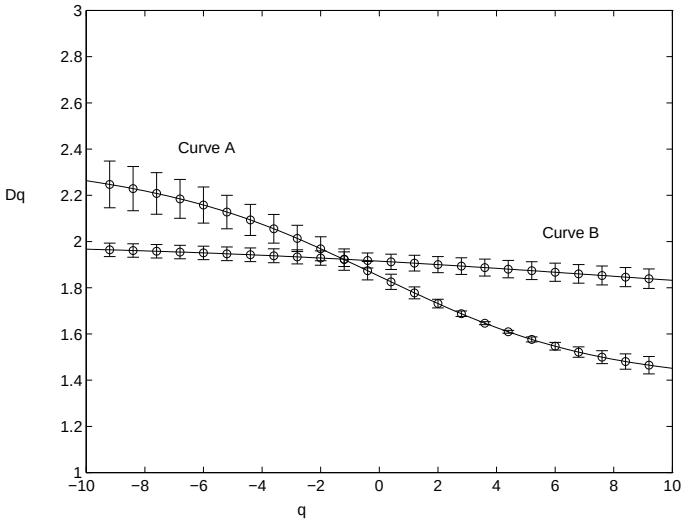

Figure 1. The spectrum of generalized dimension is shown for a subsample of a slice from Las Campanas Redshift Survey with declination $\delta = -12^{\circ}$. Curve A refers to small scales ($< 80h^{-1}$Mpc)and Curve B to large scales ($> 120h^{-1}$Mpc).

shows the generalized dimension spectrum D_q for one of the slices of the Las Campanas survey. The striking feature that is seen is that if we separately analyze the distribution for scales less than $80h^{-1}$Mpc and the one greater than $120h^{-1}$Mpc, the D_q curve shows a sharp change in its behaviour. For scales less than $80h^{-1}$Mpc the distribution behaves clearly like a multifractal. For larger scales, D_q becomes constant with q. This would have indicated a monofractal behaviour. However, since the value of D_q is nearly equal to 2 which is the same as the Euclidean dimension, this feature indicates a transition to homogeneity.

Further, we note that the difference between the two regions is more prominent if we consider the entire D_q curve rather than just the value at $q = 2$. As we have seen $q = 2$ corresponds to the correlation dimension which is related to the correlation function. This indicates that the generalized dimension brings out the transition more clearly as compared to the correlation function analysis.

Thus the fractal analysis shows

- that the universe is indeed homogeneous over very large scales.
- The transition to homogeneity occurs at a scale of around $100h^{-1}$Mpc.
- The multi-fractal analysis shows this transition much more clearly as compared with the correlation function analysis.

References

Baryshev, Y.V., Bukhmastova, Y.L., 2004, *Astr. Lett.*, **30**, 444.
Bharadwaj, S., Gupta, A.K., Seshadri, T.R., 1999, *A&A*, **351**, 405.
Borgani, S., 1995, *Phys. Rep.*, 251, 1.
Falconer, K., 1990 in Fractal Geometry: Mathematical Foundations and Applications, John Wiley.
Feder, J., 1989, in Fractals, Plenum Press.
Martinez, V. J., 1991, in Heck A., Perdang J.M., eds, Applying Fractals in Astronomy, Springer-Verlag, p. 135.
Martinez, V.J., Jones, B.J.T., 1990, *MNRAS*, **242**, 517.
Pan J., Coles P., 2000, *MNRAS*, **318**, L51.
Peebles, P.J.E., Principles of Physical Cosmology, Princeton University Press, New Jersey.
Shectman, S.A., Landy, S.D., Oemler, A., Tucker, D.L., Lin, H., Kirshner, R.P., Schechter, P.L., 1996, *ApJ*, **470**, 172.
van der Weygaert, R., Jones, B.J.T., 1992, *Phys. Lett.*, **A169**, 145.