Multifractal analysis of the galaxy distribution: Reliability of results from finite data sets

Stefano Borgani
International School for Advanced Studies, Strada Costiera 11, Trieste, Italy
and Istituto Nazionale di Fisica Nucleare, c/o Dipartimento di Fisica dell'Università, Perugia, Italy

Giuseppe Murante and Antonello Provenzale
Istituto di Cosmogeofisica del Consiglio Nazionale delle Ricerche, Corso Fiume 4, Torino, Italy

Riccardo Valdarnini
International School for Advanced Studies, Strada Costiera 11, Trieste, Italy

(Received 24 July 1992; revised manuscript received 23 December 1992)

We test the reliability of the different generalized fractal dimension estimators, when applied to point distributions with *a priori* known scaling properties. We consider the effects of varying the amount of available data and the dimensionality of the distribution. The present work is motivated by the growing interest in cosmological context to safely analyze the scale-invariant properties of the observed galaxy distribution; these results may also be of value in all physical situations where the statistical analysis of a fractal "dust" is required. We consider (a) a monofractal structure with dimension $D=1$, (b) a multifractal structure, and (c) a scale-dependent structure, behaving like a $D=1$ monofractal at small scales and an homogeneous dust at large scales. For this structure, the clustering strength and the point number density have been chosen as to be similar to those observed for the galaxy distribution. Although the different methods display different advantages and pitfalls, we find that the presently available galaxy samples can be usefully employed to trace the scaling properties generated by nonlinear clustering.

PACS number(s): 05.45.+b, 02.50.−r, 98.60.Eg

I. INTRODUCTION

In recent years, the availability of extended galaxy redshift samples has significantly improved our view of the large-scale structure of the Universe. One striking feature of large-scale clustering is the remarkable hierarchical arrangement in the distribution of observable structures: galaxies are not randomly distributed, but tend to be clumped to form clusters, while clusters form in turn even larger structures, the superclusters, involving scales of several tens of megaparsecs. This remarkable behavior led several researchers to interpret the large-scale galaxy distribution in terms of a fractal structure [1–12]. The by-now classic results on the power-law behavior of the two-point galaxy correlation functions, as well as the hierarchical behavior of the n-point functions (see, e.g., Ref. [3] for a review), indicate that the large-scale structure of the Universe possesses well-defined scaling properties, at least on scales $<10h^{-1}$ Mpc. (As usual in cosmology, h represents the value of the Hubble constant in units of 100 $\mathrm{km\,s^{-1}Mpc^{-1}}$. In general $0.5 \le h \le 1$ is considered.) The interpretation of this scaling behavior in terms of fractal clustering has recently stimulated various quantitative analyses of both the observed galaxy [5,6,9–11] and cluster distribution [12], as well as of cosmological N-body simulations [13]. To this purpose, a variety of methods have been used, which have been originally developed in the framework of statistical mechanics and dynamical systems theory for evaluating the spectrum of generalized fractal dimensions. Among these, we mention the evaluation of the generalized correlation integrals [14,15], the classic box-counting algorithm [4], the density-reconstruction procedure [16], the nearest-neighbor method [17,18], and the recent method based on the minimal spanning tree algorithm [10,19].

The various generalized dimension estimators, however, are based on different assumptions and they may be affected by different systematic errors. In addition, these methods may be sensitive in a different way to the scaling properties on different scale ranges, a fact which may be of relevance in the analysis of natural fractal sets, where the scaling behavior may be confined to a finite range of scales and where different types of fractal properties may be encountered at small and large scales. This aspect is extremely relevant in a cosmological context, where fractality of galaxy clustering is detected at small scales, while homogeneity is expected to hold at large enough scales. For the above reasons, some differences should *a priori* be expected among the results provided by the various multifractal analysis methods. The limited statistics normally encountered in the study of galaxy samples may be another source of problems. Analogously, the presence of boundary effects (related to the peculiar shapes of galaxy surveys) may potentially affect the results of the analysis.

In order to assess the reliability of the results provided by the different multifractal estimators when dealing with a finite number of data points, in this work we apply the different algorithms to fractal distributions with known scaling properties and with different statistics. We also analyze a scale-dependent monofractal distribution to explore how the different methods are sensitive to scale changes in the fractal behavior. This is particularly interesting since it turns out that a scale-dependent monofractality may sometimes be seen as a spurious multifractality. In our opinion, these tests are a necessary step in order to obtain reliable estimates of multifractal properties from galaxy data. Clearly, the results discussed here should be of value for any fractal analysis of points distributions with a finite amount of data, independent of their physical origin.

II. FRACTAL DIMENSION ESTIMATORS

A correct definition of a fractal set is "a mathematical object whose fractal (Hausdorff) dimension D is strictly larger than its topological dimension D_T." The rigorous definition of fractal dimension may be found, for example, in Mandelbrot [4]. For a fractal point set in a three-dimensional ambient space (such as the galaxy distribution), it is $D_T=0$ and $0<D\leq 3$. If $D=3$, the point set is space filling on scales larger than the mean interparticle distance. This behavior is the signature of homogeneity; in this case the fractal nature of the distribution is rather trivial. If $D<3$, then the point set does not fill the ambient space. For a fractal distribution, one has that

$$N_b(r)\propto r^{-D_0} \tag{1}$$

at small r, where $N_b(r)$ is the number of boxes with side r that are needed to cover the distribution under study. The quantity D_0 is the "box-counting dimension," which provides an estimate of the fractal dimension D.

The simplest fractal distributions are self-similar monofractals (homogeneous fractal sets), which are characterized by a single fractal dimension and by a unique scaling behavior, i.e., all moment of the probability distribution scale equivalently. More complex fractal sets are represented by the so-called multifractals [20–22]. For a multifractal set, a single fractal dimension is not sufficient to characterize the scaling properties of the distribution, and an entire spectrum of generalized fractal dimensions D_q is required. The different generalized dimension estimators have the goal of evaluating this spectrum.

The most intuitive approach to the evaluation of the generalized dimensions is based on an extension of the box-counting (BC) algorithm. For a set of N points, we define the partition function

$$B(r,q)=\frac{1}{N_b(r)}\sum_{i_b=1}^{N_b(r)}[p_{i_b}(r)]^q , \tag{2}$$

i_b being the label of the box, $p_{i_b}(r)=n_{i_b}(r)/N$, and $n_{i_b}(r)$ the number of points falling in the i_bth box. For a fractal set, at small r we expect $B(r,q)\propto r^{\tau(q)}$, where

$$\tau(q)=(q-1)D_q \tag{3}$$

and D_q is the generalized dimension of order q. The BC dimension is found for $q=0$. For a monofractal, the generalized dimensions are all equal, while multifractal sets are characterized by $D_q<D_{q'}$ when $q>q'$. According to the definition (2) of $B(q,r)$, it turns out that for $q>0$ overdense regions are mostly weighted, while negative-order dimensions deal with the scaling of the distribution inside the underdense regions. (See, for example, Ref. [22] for a technical introduction to multifractals.) Note also that the generalized fractal dimensions are rigorously defined only in the limit $r\rightarrow 0$. For "physical" fractals, however, different "effective" fractal dimensions may be associated with different scale ranges, provided that a power-law behavior of $B(r,q)$ is observed on a sufficiently large interval. In this case, the generalized dimension estimators provide information on both the scaling properties and the extent of the scaling range. In this respect, we note that every set composed by a finite number of data points has a fractal dimension $D=0$ in the limit $r\rightarrow 0$, since at very small scales the dimension of each single point is estimated.

We now briefly review the other generalized dimension estimators that will be used in this work. A useful technique is based on an extension of the correlation-integral (CI) method [14,15]. The partition function is defined as

$$Z(r,q)=\frac{1}{N}\sum_{i=1}^{N}[C_i(r)]^{q-1} , \tag{4}$$

where $NC_i(r)$ represents the number of neighbors with r from the ith point. For a fractal set, at small r the scaling $Z(r,q)\propto r^{\tau(q)}$ holds and gives the D_q dimension according to Eq. (3). The well-known correlation dimension [14] is found for $q=2$.

Another method to evaluate the spectrum of the D_q's, called the density-reconstruction (DR) method, is based on the evaluation of the partition function [16]

$$W(p,\tau)=\frac{1}{N}\sum_{i=1}^{N}[R_i(p)]^{-\tau} , \tag{5}$$

where $R_i(p)$ is the radius of the smallest sphere (centered on the ith point) containing Np points, with $2/N\leq p\leq 1$. For a fractal set, it is $W(p,\tau)\propto p^{1-q}$ at small p values. Note that in this case one obtains q (and consequently the dimension) as a function of τ. From this, the generalized dimensions may be easily obtained through Eq. (3).

The nearest-neighbor (NN) algorithm [17,18] is based on the partition function

$$G_k(n,\tau)=\frac{1}{n}\sum_{i=1}^{n}[\delta_i^{(k)}(n)]^{-\tau} , \tag{6}$$

where $\delta_i^{(k)}(n)$ is the distance of the ith particle to its kth nearest neighbor and n is the number of points in a randomly selected subsample of the distribution. In general, the evaluation of the first neighbor ($k=1$) is affected by small-scale random errors; for this reason it is preferred to use $k=3$ or $k=4$. For a fractal distribution $G_k(n,\tau)\propto n^{q-1}$, independent of the neighbor order. Again, one obtains q as a function of τ.

A last method to compute the spectrum of generalized dimensions has recently been proposed [10,19]. This is based on the evaluation of the minimal spanning tree (MST) connecting the points of the distribution. For a given point set, the MST is defined as the unique graph connecting all the points, with no closed loops and having minimal length. In this approach, a partition function is defined as

$$S(m,\tau)=\frac{1}{m}\sum_{i=1}^{m}[\lambda_i(m)]^{-\tau}\ , \tag{7}$$

where $\lambda_i(m)$ is the length of the ith link in the MST and m is the total number of links composing the MST. For a fractal set, extracting randomly selected subsamples having different number m of points, we have $S(m,\tau)\propto m^{q-1}$. Also in this case one obtains q as a function of τ.

It is important to note that there is a crucial difference between the first two methods (BC and CI) and the remaining three. In fact, the first two algorithms evaluate the partition function by *a priori* fixing the scale r. The "effective" dimension $D_q(r)$ (as given by the local logarithmic slope of the partition function) is thus a function of the physical scale r. This fact allows for disentangling the contributions of different scaling regimes at different scales, i.e., for detecting a scale-dependent fractal behavior (or, eventually, a nonscaling behavior). The other methods, however, evaluate the partition functions as functions of the probability p or of the number of points in random subsamples. All these quantities do not bear a one-to-one correspondence with the physical scale; e.g., in the DR method a given probability is associated with a broad distribution of scales, providing information on the distance scale only on average. As a consequence, the behavior of the partition function at a given value of p, n, or m mixes several contributions from different scale ranges; this may cause trouble in situations where different scaling regimes are present at different scales. In addition, the shape of the scale distribution is a function of τ, being narrower for large values of τ and much broader for negative τ's. This dependence on τ leads to weighting the various scales in a different way at different values of τ; a monofractal distribution with two scaling regimes at different scales may thus be spuriously viewed as a multifractal distribution when analyzed with these methods. An example of such a behavior is given in Sec. III C below.

III. EVALUATING THE DIMENSION SPECTRUM OF FRACTAL DISTRIBUTIONS

In order to test the properties and the pitfalls of the dimension estimators discussed above, we now apply them to the analysis of fractal structures whose scaling properties are known *a priori.* The point distributions considered here have been generated by a modification of the β model and random β model of turbulence [23,24], which have recently been proposed as simplified models of the large-scale distribution of galaxies [25–28]. Such models provide fractal point distributions through a cascading process, which in the context of turbulence modeling represents the energy transfer from large scales to small scales, where dissipation occurs. For the cosmological interpretation of these models see, e.g., Refs. [26,27]. To implement the cascading process, we start with a "parent" cube of side L, which breaks into 2^3 "child" subcubes having side $L/2$. Let f_i $(i=1,\ldots,8)$ be the fraction of the mass of the parent cube which is assigned to the ith subcube. By repeating k times this cascade iteration, we end up with 2^{3k} small cubes with side $L/2^k$, each containing a fraction of the total mass, that depends on its fragmentation history. The subsequent mass distribution can be uniquely related to the D_q spectrum of generalized dimensions. In the limit of an infinite number of iterations, it can be proven that

$$D_q=\frac{\log_2\sum_{i=1}^{8}f_i^{q-1}}{1-q}\ , \tag{8}$$

where mass conservation requires $\sum_{i=1}^{8}f_i=1$. According to Eq. (8), the number of nonvanishing f_i's determines the value of the Hausdorff dimension, while the asymptotic values $D_{-\infty}$ and $D_{+}\infty$ are fixed by the smallest and largest f_i, respectively. Once the final density field is obtained, its Monte Carlo sampling gives the required point distribution. A particularly simple case is when all the nonvanishing f_i's take the same value. In this case, Eq. (8) gives a monofractal spectrum, with the dimension value uniquely fixed by the number of nonvanishing f_i's. A homogeneous space-filling distribution is obtained when the f_i's are all equal and different from zero, so that the mass is equally distributed between all the subcubes.

Another interesting case occurs when the f_i values change with the iteration step. The corresponding structure is not self-similar, but it has different scaling properties on different scale-ranges, or a nonscaling behavior, depending upon the selected scale dependence of the f_i's.

A. Monofractal distribution

We start our analysis by considering a monofractal point distribution with $D=1$. According to Eq. (8), this can be obtained from the cascading process described above by taking only two nonvanishing f_i's, each holding $\frac{1}{2}$. To obtain a more realistic distribution, we take the number n_i of nonvanishing f_i's for each breaking object as a random variable with mean value $\langle n_i\rangle=2$; this is obtained by prescribing each object in the iteration cascade to have a probability $p=\frac{1}{4}$ of being associated with a nonvanishing f_i.

Figure 1 reports the results of the multifractal analysis of the corresponding point distribution. To check the sensitivity of the various methods to changes in the statistics, we consider both a distribution with about 18 000 points and a random subsample of 3000 points. In Fig. 1, the three different columns report the results of the methods introduced in the preceding section. The different panels in each column refer to different values of q or τ; they report the local logarithmic slope of the partition functions, as obtained by a linear least-squares fit over three adjacent values (in log-log coordinates). Here and in other following plots, solid circles refer to the entire distribution, open triangles refer to the random subsample. A meaningful value of the "effective" generalized dimension is defined by the constancy of the logarithmic slope over a sufficiently wide range of scales. A

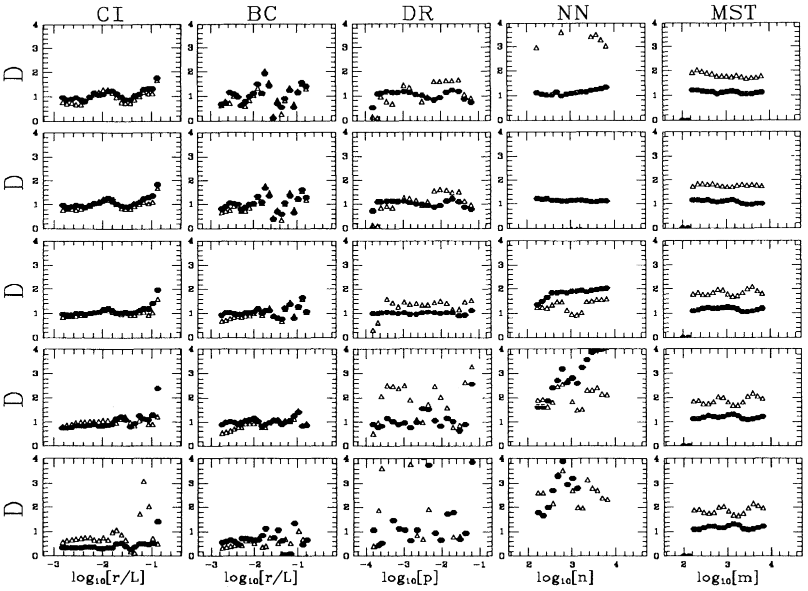

FIG. 1. The local dimension as estimated from the slope of the partition functions for a pure monofractal structure with $D=1$; the local slopes have been obtained as a running least-squares fit over three adjacent values of the partition function. Solid circles refer to the complete distribution and open triangles refer to the 3000-point random subsample. Column 1 reports the results of the CI method, column 2 refers to the BC method, column 3 to the DR method, column 4 to the NN method, and column 5 to the MST method. In columns 1 and 2 the five panels refer to $q=-2$, 0, 2, 4, and 6 from bottom to top. In column 3, the five panels refer to $\tau=-4$, -2, 0.1, 4, and 6 from bottom to top. In columns 4 and 5, the panels refer to $\tau=-6$, -4, -2, 0.1, and 2 from bottom to top.

monofractal distribution is revealed by the equality of the generalized dimensions corresponding to different values of q or τ.

Column 1 indicates that the CI method works rather well for $q \geq 0$, both for the complete distribution and for the random subsample. Note also the oscillations presented by the local dimension around the correct value. This is not a spurious artifact of the fractal algorithm, but is the consequence of the "lacunarity" (i.e., the presence of big voids with approximately periodic structure) generated by the cascading process [17]. For negative q's this method does not provide the correct results. For example, for $q=-2$ the method indicates a well-defined scaling behavior with $D_{-2}=0.4$ in the case of the complete distribution. This is due to lack of statistics and discreteness effects, which heavily affect the results at negative q's, when underdense regions with very few points are mostly weighted. Care should thus be taken in considering the results provided by this method for $q<0$. Similar behavior is evident in the results provided by the BC method (column 2), although some larger scatter of the local dimension is observed. Reliable results are obtained for $0 \leq q \leq 4$; the scatter becomes rather strong for $q>4$. This scatter is probably due to the difficulty of the BC algorithm to follow the oscillations of the local dimension due to lacunarity effects. For negative q's, the method seems to work better than the CI approach, even though a dimension less than the correct value $D=1$ is detected.

The results of the DR method (column 3) are particularly interesting. In fact, the correct dimension $D=1$ is estimated for $\tau>0$ in the case of the entire distribution; however, for the random subsample there is a tendency towards estimating a larger dimension; for example, the estimated dimension is $D \approx 1.4$ for $\tau=0.1$ for the 3000 points sample. This trend becomes rather dramatic for negative τ's. For $\tau<0$ an average dimension $D \approx 1$ is correctly estimated from the entire distribution (even though with a noticeable scattering of the local slope); however, for the 3000 points sample the dimension estimates are much larger (e.g., $D_\tau \approx 2$ for $\tau=-2$). Analysis on much larger distributions (e.g., a fractal with $D=1$ and 100 000 points) has revealed that the DR method provides extremely reliable results for large data sets. However, the results in Fig. 1 show that this method is very sensitive to the problem of limited statistics, especially for values of $\tau \leq 0$: the resulting variation of the dimension with τ simulates some spurious multifractality.

Column 4 reports the results for the NN method when the fourth-order neighbor is considered for $\tau \geq 0$ and the first-order neighbor is used for $\tau<0$. The analysis has been repeated for the first four orders of neighbors, for all values of τ. The results for all these neighbor orders are very similar; the above choice minimizes random scatter and fluctuations in the logarithmic slope of the partition function. In general, for the entire distribution this method provides the correct results for $\tau \geq 0$; the results for the random subsample are not reliable even for positive τ's. For negative values of τ, the NN method does not give the correct results, for neither the entire distribution or the random subsample. The discreteness problems present for $\tau<0$ are sometimes translated by this algorithm into a wild scattering of the local slope of the partition function. On the other hand, note that the method provides a well-defined, but wrong, estimate $D \approx 2$ for $\tau=-2$ for the entire distribution. This behavior is rather critical since the flatness of the local dimension may lead to incorrect conclusions.

Column 5 shows the results for the MST method. This method provides the correct result $D=1$ for all values of τ, in the case of the entire distribution. For the same values of τ, the results for the random subsample provide $D \approx 2$. Note that the local logarithmic slope of the partition function is apparently well behaved also for the random subsample; however, the convergence is forced to a wrong value of the dimension. To optimize the performance of the MST method, it has been suggested [19] to eliminate the very small (for $\tau > 0$) and the very long (for $\tau < 0$) edge links from the construction of the MST. These edge links are in fact likely to introduce small- and large-scale noise, respectively. Progressively cutting the tails of the edge-links distribution induces convergence to a well-defined local logarithmic slope of the partition function. For the case studied here, this procedure furnishes the correct results, since the partition function has either no scaling behavior (for the wrong edge-links cuts) on a local logarithmic slope giving $D=1$. As a conclusion, we observe that the MST method gives a correct answer for all values of τ, when a distribution with a sufficiently large statistics is analyzed. In general, this method should, however, be used with with great care on distributions with limited statistics in order to avoid apparent convergence of the local slope and consequent spurious estimates of the dimensions.

B. Multifractal distribution

As a second step, we consider the analysis for a multifractal distribution with dimension spectrum fixed according to Eq. (8). For the purpose of the present study we have chosen the f_i's such that $D_\infty = 0.8$, $D_0 = 2$, and $D_{-\infty} = 3$. The field produced by the random β model has been sampled with a total of 50 000 Monte Carlo points. Figure 2 reports the local dimensions estimated by the various partition functions.

Column 1 reports the results for the CI method. Note the growth of the local dimension at small scales, due to the presence of small-scale Gaussian noise generated by the Monte Carlo sampling. This is more evident at larger values of q; in fact, in the regions of high density there is a larger number of points, which are uniformly random distributed inside the same box. In this case, the average distance between points in high-density regions is less than the minimum scale generated by the random β model; at very small scales the distribution has dimension three. At larger scales, the correct dimension is determined for values of $q > 0$, even though the local scope displays non-negligible functuations. For $q \leq 0$ the results provided by the CI method do not reproduce the correct dimension spectrum, as already noted in the study of the monofractal dust with $D=1$. In the present case, the D_0 dimension is not correctly evaluated by the CI approach.

The results of the BC method are reported in column 2. The local slope of the partition function is extremely scattered in this case, while discreteness effects do not allow the evaluation of the dimensions for $q \leq 0$. For positive q's, the average value of the logarithmic slope of the partition function provide an approximate estimate of the corresponding generalized dimension.

Column 3 reports the results for the DR method. For $\tau \gg 0$, this method is able to separate the (small) scales, where random noise dominates, from the (larger) scales, where the distribution is fractal. For example, in the case of the entire distribution, for $\tau = 6$ there is a large plateau with $D_\tau \approx 2.5$ for $p < 0.01$; the dimension estimates assume their correct value for $p > 0.01$. For the random subsample this effect is even more evident; the dimension estimate is $D_\tau \approx 3$ for large τ's and $p < 0.01$. For $\tau \leq 0$, the dimension estimates for the entire distribution are close to the theoretical values, even though they tend to remain below the correct values, especially for $\tau \ll 0$, because of discreteness effects. For $\tau \leq 0$ the results for the random subsample provide dimension estimates that are larger than the correct ones. In general, the D_q estimates provided by the DR method are rather reliable once the local slope is evaluated over a p range where small-scale noise is absent. This is clearly shown in Fig. 3, which reports the dimension estimates obtained with this method together with the theoretical dimensions [as obtained by Eq. (8)]. A good agreement between the two spectra is obtained.

The NN method, shown in column 4 of Fig. 2, provides a correct evaluation of the multifractal spectrum

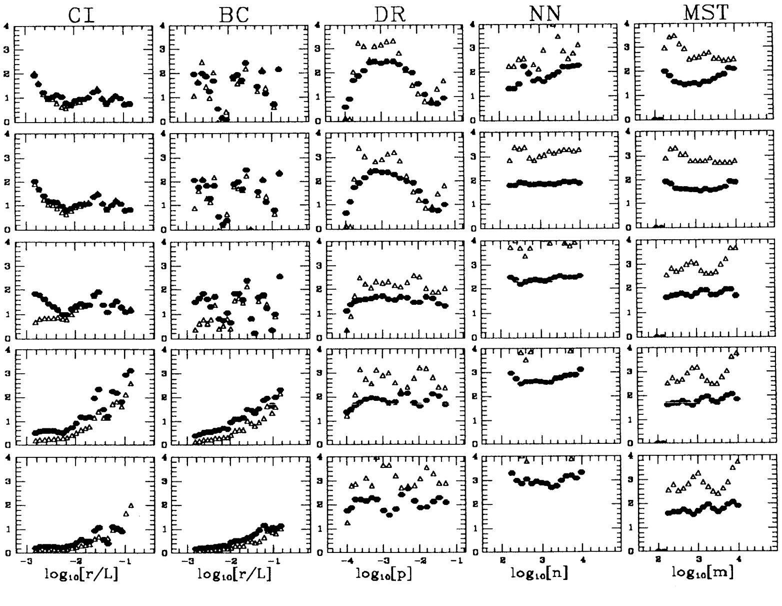

FIG. 2. The same as in Fig. 1, but for the multifractal structure discussed in the text.

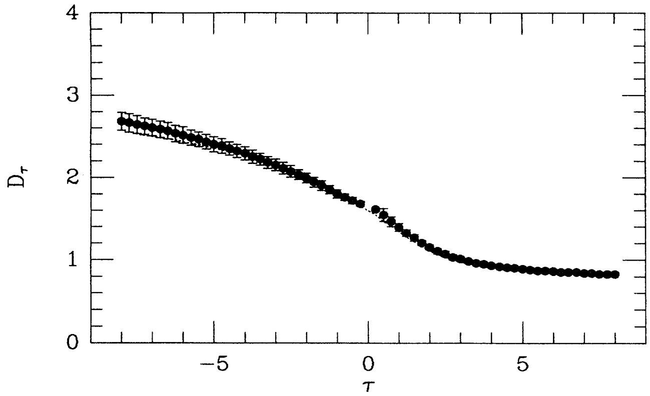

FIG. 3. Spectrum of generalized dimensions D_τ vs τ for the multifractal structure analyzed in Fig. 2. The solid line indicates the theoretical values of the dimension while solid circles indicate the dimension estimates obtained by the DR method.

for $\tau \leq 0$, if the large sample is considered. The analysis of the random subsample does not give any stable result. For positive values of τ, the analysis of the entire distribution reveals the small-scale random noise (corresponding in this case to large values of n), providing also an approximate estimate of the correct generalized dimension at small values of n. The results of the MST approach are reported in column 5 of Fig. 2. This method provides a correct dimension estimate for $\tau = -2$, corresponding in this case to $q = 0$, for the entire distribution. The dimension estimates obtained for the other values of τ do not correctly reproduce the theoretical values of the generalized dimensions. For $\tau > -2$, the dimensions remain larger than the theoretical values, while for $\tau < -2$ the dimensions remain below the correct values. In general, this method displays some tendency towards providing dimension estimates about $D \approx 2$ for this multifractal distribution. For the random subsample, the MST method does not provide reliable results. A dimension $D_\tau \approx 3$ is in fact detected for the random subsample; confirming the high sensitivity of the MST method to statistics.

For the MST method, the choice of different cuts to the edge-link distribution does not improve the evaluation of the multifractal spectrum. Figure 4 reports the local logarithmic slopes of the MST partition functions for different cuts in the edge-link distribution. The various columns correspond to different edge link cuts, the values of τ are the same as already considered in Fig. 2. Contrary to what happens in the analysis of a pure monofractal distribution, where either convergence to the correct value of the dimension or a huge scattering of the local slope is observed, for the multifractal dust the slopes appear to converge rather clearly to an approximately constant value for various choices of the edge-link cuts. The spurious estimates provided by the MST approach are not an artifact of a particular choice of the cut on the distribution; rather, they seem to be inherent in this method.

C. Scale-dependent distribution

In most natural systems, the fractal behavior does not extend over arbitrarily large-scale ranges; instead, it is observed only on a finite scaling regime. In a cosmological context, this is just the case for the large-scale distribution of galaxies. Observational evidences indicate that the multifractal properties of the galaxy distribution are confined to small scales, while $D_q \approx 3$ at scales larger than an appropriate homogeneity threshold. Such a behavior is consistent with the view that the fractal properties are built by the process of nonlinear gravitational clustering [13]. According to the results of N-body simulations of cosmological gravitational clustering, an initially homogeneous point distribution is transformed by the gravitational evolution into a fractal dust; since the gravitational clustering starts from the small scales, at every finite time there is the simultaneous presence of an evolved mul-

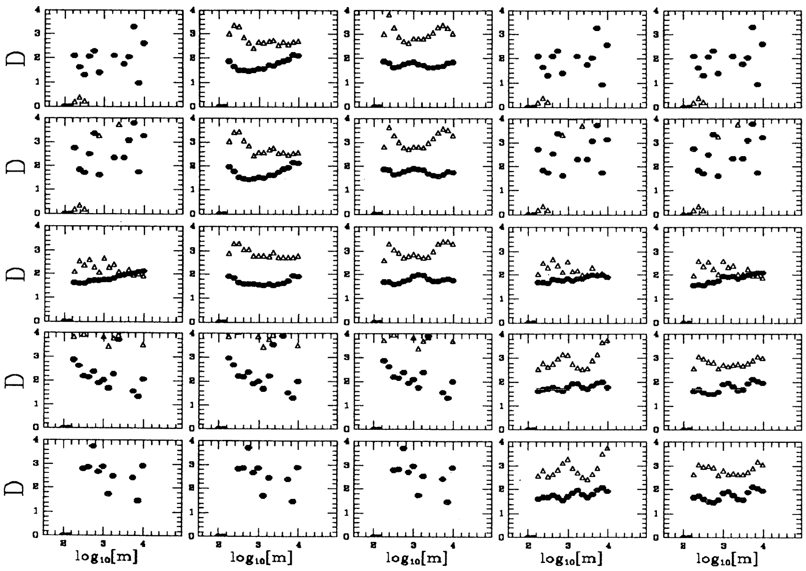

FIG. 4. The local dimension from the MST partition function for the multifractal structure and for different edge-link cuts; the local slopes have been obtained as a running least-squares fit over three adjacent values of the partition function. Solid circles refer to the complete distribution and open triangles refer to the 3000-point random subsample. Column 1 reports the results when no edge links are cut. Columns 2 and 3 report the results obtained by cutting edge links shorter than $0.1\langle L \rangle$ and $0.3\langle L \rangle$, respectively; here $\langle L \rangle$ is the mean edge-link length. Columns 4 and 5 report the results obtained by cutting edge links larger than $2\langle L \rangle$ and $4\langle L \rangle$, respectively.

tifractal distribution at small scales and of an homogeneous distribution at large scales. In this case, it is important to verify the behavior of the various analysis methods, especially of those algorithms which mix different scale ranges in the evaluation of the partition function.

To approach this problem, here we consider a scale-dependent monofractal distribution which is characterized by $D=3$ for scales larger than an homogeneity scale L_h and by $D=1$ for scales smaller than L_h. This distribution is obtained by the cascading process previously discussed, by an appropriate choice of the f_i parameters in the two different scaling ranges. The homogeneity scale is chosen to be $\frac{1}{4}$ of the size L of the simulation box; there is a total of 30 000 points in the distribution. A comparison with the galaxy distribution is made possible by requiring the number density of the simulated distribution to be approximately equal to the average number density $\langle n \rangle$ of bright galaxies, $\langle n \rangle = 0.01h^{-1}$ galaxy Mpc^3. Such a density is consistent, e.g., with the average galaxy density obtained in the CfA II [29] redshift sample. The density indicated above gives $L=140h^{-1}$ Mpc and $L_h=35h^{-1}$ Mpc in physical units. As an example of undersampling, we also analyze a 3000-point random subsample of the complete distribution.

Figure 5 shows the results of the multifractal analysis of the entire scale-dependent monofractal distribution (solid circles) and of the random subsample (open triangles). For positive q's, both CI and BC methods provide extremely reliable results for the complete distribution, indicating both the correct value of the dimension at small scales ($D=1$) and the transition to homogeneity above L_h. For $q=0$, the BC method gives a correct estimate of the dimension, while the CI approach provides a slight underestimate of the dimension. The improved reliability of the BC method with respect to the scale-free $D=1$ structure is due to the fact that large-scale homogeneity fills the voids, thus suppressing the presence of lacunarity. This is also apparent from the remarkable stability of the local dimensions revealed by the CI method.

As usual, none of these methods is able to estimate at small scales the generalized dimensions for $q<0$, due to discreteness effects, while detecting the large-scale homogeneity. Note that the CI method gives an apparently stable (but incorrect) estimate $D_q \approx 0.5$ for $q=-2$. For the random subsample, neither the CI nor BC method provides reliable results at small scales. However, both methods still detect the transition to large-scale homogeneity. This behavior is generated by the fact that many points are now found at large mutual separations, due to the imposed large-scale homogeneity. In the case of the random subsample, the statistics is thus not sufficient to correctly sample the fractal behavior at scales smaller than L_h. Analogously, for the complete distribution this effect leads to the presence of discreteness effects at scales slightly larger than those detected for the pure $D=1$ distribution, even though the total number of points in the scale-dependent distribution is larger.

The results of the DR method are reported in column 3. For $\tau \geq -2$, this method provides a reliable estimate of the fractal dimension and of the transition to homogeneity in the case of the complete distribution. For the random subsample, the results are not correct and they provide spurious estimates of the fractal behavior. As discussed in Sec. II, an important characteristic of this method is that it mixes different scale ranges in the evaluation of the partition function at a given value of p. This mixing becomes more evident as the value of τ decreases; for $\tau \leq -4$ the results can hardly be interpreted, due to a strong mixing between the small scales (where $D=1$) and the large scales (where $D=3$). Care has thus to be taken when using this method for evaluating the negative τ dimensions on scale-dependent fractal sets.

Columns 4 and 5 of Fig. 4 report the results of the NN approach and of the MST method, respectively. The random subsample provides extremely scattered and unsta-

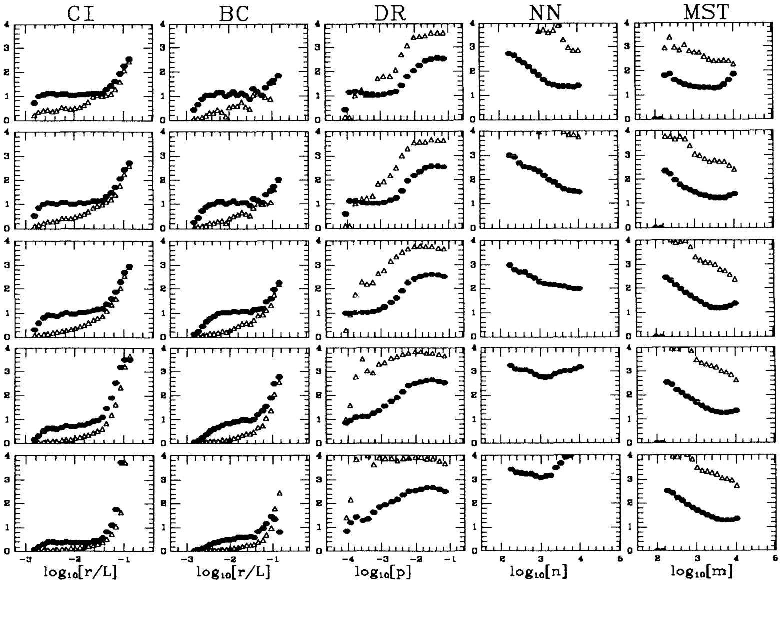

FIG. 5. The same as in Fig. 1, but for the scale-dependent fractal structure discussed in the text.

ble results when analyzed with these two methods. For positive τ's, the NN method gives somewhat reliable results for the complete distribution, with the caveat that scale mixing tends to obscure the true scale dependence of the fractal dimension for moderate values of τ. The best scale separation is obtained here, as for the previous method, for large values of τ. For $\tau<0$ scale mixing becomes dramatic (even worse than for the DR method); for example, a small-scale (large n) dimension $D_\tau \approx 2$ is evaluated for $\tau=-2$, and $D_\tau \approx 3$ for $\tau=-4$, suggesting (erroneously) the presence of a multifractal distribution. The use of this method on scale-dependent fractal sets can thus spuriously transform the presence of two scaling regimes at different scales into an apparent multifractality, since the effects of the scale mixing are different for different values of τ. The results provided by the MST are quite stable along the whole sequence of τ values. For the complete distribution, it always detects the correct values, $D=1$ and $D=3$, holding at small and large scales, respectively. However, because of scale mixing, only a smooth transition between these two values is detected, without any evidence of scale invariance over a finite interval. Again, for the smaller sample the limited statistics heavily affects the dimension estimate.

In order to check in detail the effects of undersampling, we now consider five scale-dependent distributions with different statistics, characterized by a total number of points $N=3000, 8000, 12\,000, 19\,000$, and $30\,000$, respectively. These subsets are randomly drawn from an original distribution of 60 000 points. By power-law-fitting the slope of the partition functions up to the homogeneity scale L_h, we verify the reliability of the dimension estimates for various values of q and τ. The results are plotted in Fig. 6 for the BC (a) and the DR (b) methods. The analysis with the GP method gives results which are similar to those of the BC algorithm. This is justified on the basis of the plots of the local dimensions reported in Fig. 5. As for the MST and NN methods, we do not report any result, given their limited reliability.

In Fig. 6(a) we plot the D_q dimensions (for the same q values as in Fig. 5, i.e., $q=-2,0,2,4,6$) versus the number of points N. The error bars are 3σ uncertainties arising from the log-log linear regression. For the highest statistics, the dimensions converge to the correct value $D_q=1$ at positive q's while the estimate is affected by discreteness effects at low q's. The situation becomes worse and worse when poor distributions are considered; the D_q values progressively decrease with N. This effect can be easily understood by a visual inspection of Fig. 5; reducing N restricts the scale range where D_q takes flat, going to zero at very small scales. This shows the importance of plotting the local dimensions in order to verify the width of the scale-invariance range before proceeding with a crude power-law fit of the partition function over the scales where self-similarity is expected.

Figure 6(b) reports the generalized dimensions for the DR method for $\tau=-6,-4,-2,0.1,2$. In this case, the W partition function of Eq. (5) depends on the probability measure p, instead of on the scale r. For this reason, a suitable prescription must be devised in order to associate the proper p value to the homogeneity scale L_h. From Eq. (5), it is easy to recognize that $W(p,\tau=-1)$ represents the value of the average radius associated to p. Accordingly, we fit the DR partition function up to the probability p^*, such that $W(p^*,\tau=-1)=L_h$. It is, however, clear that such a procedure is not rigorously correct, since no one-to-one correspondence exists between p and r values. As a consequence, some scale mixing always appear, whose amount increases as lower τ's are considered (see Fig. 5). As in the case of the BC method, a reliable estimate of the fractal dimension is attained only for the richest distribution. However, even in this case, D_τ is systematically overestimated, mostly at

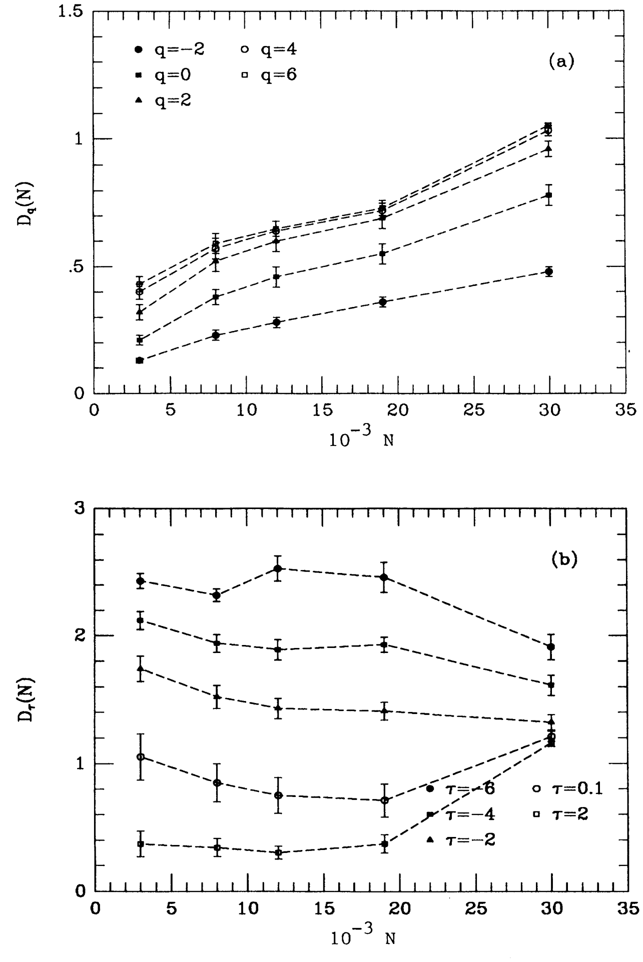

FIG. 6. Estimates of the fractal dimension for the scale-dependent distribution, over the scale range where $D=1$ is expected, as a function of the number of points used to trace the underlying structure. Different symbols refer to different multifractal orders. The plotted error bars are 3σ uncertainties arising from the log-log linear regression of the partition functions. Panel (a) is for the BC method. In this case, the dimension estimates are on the scales below the homogeneity scale L_h. Panel (b) is for the DR method. Here, the dimension estimates are realized below the probability p^*, which is associated to L_h according to $W(p^*,\tau=-1)=L_h$.

negative τ. This effect is entirely due to scale mixing; even at small p value, the partition functions take some contribution from the large scales, where homogeneity ($D=3$) holds. The scale-dependence generates a spurious multifractal behavior, which we recall to be absent by construction in the point distribution under study. Once more, a close investigation of the local dimensions is strongly suggested in order to avoid incorrect inferences. The situation becomes even worse for lower statistics; small-scale discreteness originates an underestimate of the dimension for $\tau > 0$, and, vice versa, an overestimate for $\tau < 0$.

IV. DISCUSSION AND CONCLUSIONS

In this paper we have discussed the reliability of the results provided by different fractal dimension estimators when applied to three-dimensional fractal dusts with finite statistics. In particular, we have considered structures with different fractal properties and with a different number of points. We have analyzed three different structures: a monofractal with dimension $D=1$, a multifractal, and a scale-dependent structure with $D=1$ at small scales and homogeneity ($D=3$) at large scales. The present analysis has been motivated by the growing interest in the quantitative determination of the scaling properties of the galaxy distribution [1–12] and in the study of the fractal nature associated with nonlinear gravitational dynamics [13]. It is clear, however, that the results discussed here are of more general interest, being relevant to any statistical analysis of experimental or numerical point distributions, in many different physical contexts.

The main results of this work can be summarized as follows.

(a) The box-counting and the correlation-integral methods are in general quite reliable to estimate positive-order dimensions, while they suffer for discreteness effects for negative q's, where underdense regions are mostly weighted in the computation of the partition function. The stability of these methods, when the number of points in the sample is decreased, depends on the dimensionality of the structure; fractals having a lower dimension require a smaller number of points to be adequately sampled. A further advantage of these methods is also that they fix *a priori* the physical length scale where the dimension is estimated. This aspect is of particular relevance in the analysis of scale-dependent structures. In this case, the BC and CI methods allow one to safely detect the presence of a characteristic scale in the distribution, where the dimensionality sharply changes.

(b) The DR method is rather good in estimating both positive- and negative-order dimensions. For this reason, this method is particularly suited to follow the whole spectrum of dimensions in a multifractal structure. This is clearly seen in Fig. 3, where the DR method measures D_q values, which are always remarkably similar to the true values expected on the grounds of Eq. (8). A possible drawback of this method lies in the fact that, differently from the BC and CI algorithms, each probability value does not correspond to a unique choice of the physical scale; instead, scale mixing may occur. Clearly, this could represent a potential problem when dealing with scale-dependent structures. From Fig. 5, one sees that the DR method is able to disentangle the different scaling regimes for positive τ's where scale mixing is less dramatic, while the results are much less reliable for negative τ's. Finally, this method severely suffers for lack of statistics. For poor samples, as shown in Fig. 1, the DR method tends to overestimate the dimension.

(c) As far as the MST and NN methods are concerned, they appear to give the less reliable answers. The only case in which they have been shown to be acceptable are for the $D=1$ structure with 18 000 points. This suggests that such methods are efficient only when a very high sampling rate is allowed. In the analysis of the scale-dependent structure, these methods displayed a very strong scale mixing. As a consequence, the local dimension never flattens at an approximately constant value, although it ranges between the correct values, $D=1$ and 3 at small and large scales, respectively. These results suggest that some care must be payed when using the MST and NN methods to analyze the multifractal spectrum of the galaxy distribution.

(d) A general indication provided by the results discussed here concerns the relevance of using the local dimensions in order to verify the existence and the extension of a self-similarity scale range. In fact, discreteness effects always put a lower bound to the scales where scale-invariance can be safely detected. Moreover, when different fractal properties are expected at different scale ranges (such as for the galaxy distribution in cosmological context), the presence of scale mixing in some fractal algorithms can produce dimension estimates, which detaches from the correct values by an amount depending on the multifractal order.

From the results of the present work, some general conclusions on the multifractal analysis of the galaxy distribution can be drawn. The scale-dependent structure considered here has a number of points and a homogeneity scale which are similar to those encountered in the analysis of real galaxy samples. The plots of Figs. 5 and 6 show than an excessive sparse sampling (as in the case of the random subsample) does not allow one to trace adequately the scaling properties where the clustering is nonlinear. However, the results obtained from the entire distribution indicate that the multifractal analysis methods discussed here should provide reliable results, when appropriately employed. Our conclusion is thus that complete and extended redshift samples, such as the emerging CfA II galaxy redshift survey [29], have enough statistics to trace the fractal and scaling properties associated with gravitational dynamics. A crucial requisite, however, is the knowledge of the behavior and of the pitfalls of the various analysis methods; analogously, the simultaneous usage of several different analysis methods is an important ingredient of a reliable analysis. Both these issues have been thoroughly discussed here.

[1] M. S. Soneira and P. J. E. Peebles, Astrophys. J. **83**, 845 (1978).

[2] G. Efstathiou, S. M. Fall, and G. Hogan, Mon. Not. R. Astron. Soc. **189**, 203 (1979).

[3] P. J. E. Peebles, *The Large Scale Structure of the Universe* (Princeton University Press, Princeton, NJ, 1980).

[4] B. B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman, San Francisco, 1982).

[5] L. Pietronero, Physica A **144**, 257 (1987).

[6] B. J. T. Jones, V. J. Martinez, E. Saar, and J. Einasto, Astrophys. J. Lett. **332**, L1 (1988).

[7] R. Balian and R. Schaeffer, Astron. Astrophys. **220**, 1 (1989).

[8] S. Borgani, Mon. Not. R. Astron. Soc. **260**, 537 (1992).

[9] V. J. Martinez and B. J. T. Jones, Mon. Not. R. Astron. Soc. **242**, 517 (1990).

[10] V. J. Martinez, B. J. T. Jones, R. Dominguez-Tenreiro, and R. van de Weygaert, Astrophys. J. **357**, 50 (1990).

[11] L. Guzzo, A. Iovino, G. Chincarini, R. Giovannelli, and M. Haynes, Astrophys. J. Lett. **382**, L5 (1991).

[12] S. Borgani, M. Plionis, and R. Valdarnini, Astrophys. J. **404**, 21 (1993).

[13] R. Valdarnini, S. Borgani, and A. Provenzale, Astrophys. J. **394**, 422 (1992).

[14] P. Grassberger and I. Procaccia, Phys. Rev. Lett. **50**, 346 (1983).

[15] G. Paladin and A. Vulpiani, Lett. Nuovo Cimento **41**, 82 (1984).

[16] P. Grassberger, R. Badii, and A. Politi, J. Stat. Phys. **51**, 135 (1988).

[17] R. Badii and A. Politi, Phys. Rev. Lett. **52**, 1661 (1984).

[18] R. Badii and A. Politi, J. Stat. Phys. **40**, 725 (1985).

[19] R. van de Weygaert, B. J. T. Jones, and V. J. Martinez, Phys. Lett. A **169**, 145 (1992).

[20] R. Benzi, G. Paladin, G. Parisi, and A. Vulpiani, J. Phys. A **17**, 3521 (1984).

[21] T. C. Halsey, M. H. Jensen, L. P. Kadanoff, I. Procaccia, and B. I. Shraiman, Phys. Rev. A **33**, 1141 (1986).

[22] G. Paladin and A. Vulpiani, Phys. Rep. **156**, 147 (1987).

[23] U. Frisch, P. L. Sulem, and M. Nelkin, J. Fluid Mech. **87**, 719 (1978).

[24] U. Frisch and G. Parisi, in *Turbulence and Predictability in Geophysical Fluid Dynamics and Climatology,* edited by R. Benzi, G. Parisi, and A. Sutera (North-Holland, Amsterdam, 1985).

[25] C. Castagnoli and A. Provenzale, Astron. Astrophys. **246**, 634 (1991).

[26] A. Provenzale, in *Applying Fractals in Astronomy,* edited by A. Heck and J. Perdang (Springer, Berlin, 1992).

[27] A. Provenzale, P. Galeotti, G. Murante, and B. Villone, Astrophys. J. **401**, 455 (1992).

[28] B. J. T. Jones, P. Coles, and V. J. Martinez, Mon. Not. R. Astron. Soc. **256**, 146 (1992).

[29] V. de Lapparent, M. J. Geller, and J. P. Huchra, Astrophys. J. **369**, 273 (1991).