Pseudofractals and the box counting algorithm

2001 J. Phys. A: Math. Gen. 34 7933

(http://iopscience.iop.org/0305-4470/34/39/302)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **34** (2001) 7933–7940 PII: S0305-4470(01)24172-9

Pseudofractals and the box counting algorithm

Andrzej Z Górski

Institute of Nuclear Physics, Radzikowskiego 152, 31–342 Kraków, Poland

E-mail: Andrzej.Gorski@ifj.edu.pl

Received 17 April 2001, in final form 17 July 2001
Published 21 September 2001
Online at stacks.iop.org/JPhysA/34/7933

Abstract

We show that, for sets with the Hausdorff–Besicovitch dimension equal to zero, the box counting (BC) algorithm commonly used to calculate the Renyi exponents (d_q) can exhibit perfect scaling suggesting non-zero d_q. The properties of these pathological sets (*pseudofractals*) are investigated. Numerical, as well as analytical, estimates for d_q are obtained. A simple indicator is given to distinguish pseudofractals and fractals in practical applications of the BC method. Histograms made of pseudofractal sets are shown to have Pareto tails.

PACS numbers: 47.52.+j, 02.60.Gf, 05.45.Df

1. Introduction

The notion of fractals was introduced in the 1970s by Mandelbrot and soon became very fashionable. In the mathematical sense, a set is called a fractal (set) when its Hausdorff–Besicovitch dimension (d_{HB}) is greater than its topological dimension (d_{T}) [1]. Since fractality is strictly related to the physically important self-similarity (self-affinity) scaling symmetries and the renormalization group, it is widely used in physics on all scales: ranging from particle physics [2] to astrophysics [3], and in various other areas such as solid-state physics [4] or econophysics [5].

However, in contrast to fractal sets constructed by mathematicians, such as the famous triadic Cantor set (1883), for physically interesting cases, the algorithms used to construct corresponding data sets are usually unknown and it is very difficult (or almost impossible) to calculate their Hausdorff–Besicovitch dimension, Renyi exponents etc in a mathematically rigorous way. Instead, one considers a zero-dimensional (finite number) subset of data points and applies a standard numerical algorithm, such as the box counting (BC) algorithm or its derivatives, which gives the well known log–log plot. A good linear fit is assumed to be equivalent to the calculation of the corresponding fractal dimensions. Apart from the fact that the above fit is to some extent arbitrary (see, for example, [6]) and there is no good method to calculate 'error bars', it will be shown in the following sections that even a perfect fit can be misleading.

In fact, there are many different mathematical definitions of the fractal (capacity) dimension that can give different results when applied to the fractal set. These were originally introduced to physics to characterize strange attractors of dynamical systems [7]. The BC capacity dimension is closest to the dimension introduced by Kolmogorov [8].

Here, we limit our discussion to the BC method that is the basic paradigm for a practical computation of the generalized Renyi exponents, d_q, defined by [9, 10]

$$d_q = \frac{1}{1-q} \lim_{N\to\infty} \frac{\ln \sum_i p_i^q(N)}{\ln N} \equiv \lim_{N\to\infty} \frac{\ln Y(N)}{\ln N} \tag{1}$$

where N is the total number of 'boxes' (bins), and p_i is the part of the 'mass' (i.e. the fraction of all points) contained in the ith box. In this paper we also deal with sets that are not fractals, namely the discrete (point) sets. For these sets one defines

$$p_i(N) = \frac{n_i(N)}{n_{\text{tot}}} \tag{2}$$

where $n_i(N)$ is the number of data points ('mass') in the ith box for a given subdivision (partition) N and n_{tot} is the total number of data points ('mass') contained in all boxes. When $q = 0$ (capacity dimension) equation (1) becomes

$$d_0 = \lim_{N\to\infty} \frac{\ln M(N)}{\ln N} \tag{3}$$

where $M(N)$ denotes just the number of non-empty boxes. In this case, the number of data points in particular boxes is irrelevant and this singles out the value $q = 0$. This is the reason why the BC method gives a unique result for d_0 (see section 2). d_q is determined from the log–log plot of $\log Y(N)$ versus $\log N$ with $N = 2^0, \ldots, 2^k$, usually with $k \simeq 10$–30.

Since in practical computations with the BC and derivative methods one always deals with a finite number of data points, we limit our analysis to discrete sets. In the following section we obtain an analytic expression for d_0 with the set defined by [11, 12]

$$x_n = \frac{1}{n^a} \qquad n = 1, 2, \ldots \quad a > 0. \tag{4}$$

The same method is also applied for general discrete sets with an accumulation point, as well as for divergent series. In section 3, the BC algorithm is applied to calculate the Renyi exponents with $q \neq 0$ for equation (4). Excellent scaling (linear fit) has been found in full agreement with analytical estimates, in spite of the fact that the set (4) is not a fractal and its Hausdorff–Besicovitch dimension equals zero. Also, it is shown that the standard BC method leads to a violation of the Hentschel–Procaccia (HP) inequality [10]. A modification of the standard BC method which preserves the HP inequality is analysed as well. Our results can be generalized to sets with an arbitrary number of accumulation points. In section 4 it is shown that pseudofractals generate histograms with fat tails, in contrast to fractals. A final discussion is given in the last section.

2. Capacity dimension of pseudofractals

Clearly, the discrete and countable set (4) is not a fractal and it has a zero Hausdorff–Besicovitch dimension. However, as has been demonstrated using the dimension function [11] or by direct application of the BC method [12], numerical computation must give the following analytic result:

$$d_0 = \frac{1}{1+a}. \tag{5}$$

As the method of the analytical estimates of [12] and its generalization will be used throughout the paper, we now describe this briefly. Assuming the unit size of the whole set, the size of a single bin is $1/N$. Denoting the number of bins by N_{sngl} (and the number of corresponding data points by $n_{\mathrm{sngl}} = N_{\mathrm{sngl}}$) with one, and only one, data point inside, one can easily calculate from equation (4)

$$n_{\mathrm{sngl}} = N_{\mathrm{sngl}} \sim N^{\frac{1}{1+a}}. \tag{6}$$

Since we have a logarithm in (1) and (3), and the limit $N \to \infty$, the constant pre-factor can be neglected. The remaining data points (n_{r}) are closer to each other than the bin size. Hence, all those bins are non-empty. The number of such bins ($N_{\mathrm{r}} < n_{\mathrm{r}}$) is equal to the distance of the point $x_{n_{\mathrm{sngl}}}$ from the accumulation point ($x_\infty = 0$) divided by the bin size ($1/N$). This gives the estimate

$$N_{\mathrm{r}} \sim N^{\frac{1}{a}} \tag{7}$$

and in the limit $N \to \infty$

$$M(N) \sim N_{\mathrm{sngl}} + N_{\mathrm{r}} \sim N^{\frac{1}{1+a}} + N^{\frac{1}{a}} \sim N^{\frac{1}{1+a}} \tag{8}$$

which implies the result (5).

From the above proof it is clear that the exponent d_0 depends on the rate of change of the distances between neighbouring points ('level spacing') with respect to the length of the whole interval or, in other words, on the speed of the convergence of data points to the accumulation point. This enables us to generalize the above result. Also, one can consider divergent sets ($x_n \to \infty$ for $n \to \infty$) by rescaling them to the unit interval. To this end, we define the convergence rate $\Delta x(n)$ by

$$\Delta x(n) = \begin{cases} |x_n - x_{n+1}|/|x_1 - x_\infty| & \text{for} \quad |x_\infty| < \infty \\ |x_n - x_{n+1}|/|x_n| & \text{for} \quad |x_\infty| = \infty. \end{cases} \tag{9}$$

This gives the following general formula for the exponent d_0

$$d_0 = \min\left\{ \lim_{n\to\infty} \frac{-\ln n}{\ln \Delta x(n)},\ 1 \right\}. \tag{10}$$

In particular for equation (4), one obtains $\Delta x(n) \sim 1/n^a$ that leads to formula (5). For slowly converging series, such as $1/\ln n$, one has $d_0 = 1$, while for strong convergence (e.g. $x_n = \mathrm{e}^{-an}$) one has $d_0 = 0$. On the other hand for all diverging series (such as n^a, e^{+an} or $\ln n$) one always has $d_0 = 1$. Intuitively, slowly converging series seem to be uniformly distributed, while those exponentially converging seem to be concentrated at the accumulation point (zero dimensional). Hence, from this point of view, the series with inverse power asymptotic are the only non-trivial ones.

The above results can be verified numerically by applying the BC method. The results are displayed in figure 1, where the straight lines correspond to the theoretical predictions. Actually, for the 10^4 data points, one can already see an excellent linear scaling in the log–log plot through more than a dozen binary orders of magnitude—well above what is usually demanded in practical applications. In addition, the results are in perfect agreement with formula (10): $d_0 = 0.50$, 0.66 and 0.33 for $a = 1$, 0.5 and 2, respectively, while for divergent series, $\sqrt{n}$ and n^2 (crosses and circles respectively), one obtains $d_0 = 1.0$.

3. Pseudofractals and generalized Renyi exponents

For $q \neq 0$ analytical estimates are ambiguous as we have to deal with the double limit: $\lim_{N\to\infty} \lim_{n_{\mathrm{tot}}\to\infty}$, because the probabilities ($p_i = p_i(n_i, n_{\mathrm{tot}}, N)$) do depend on both N and

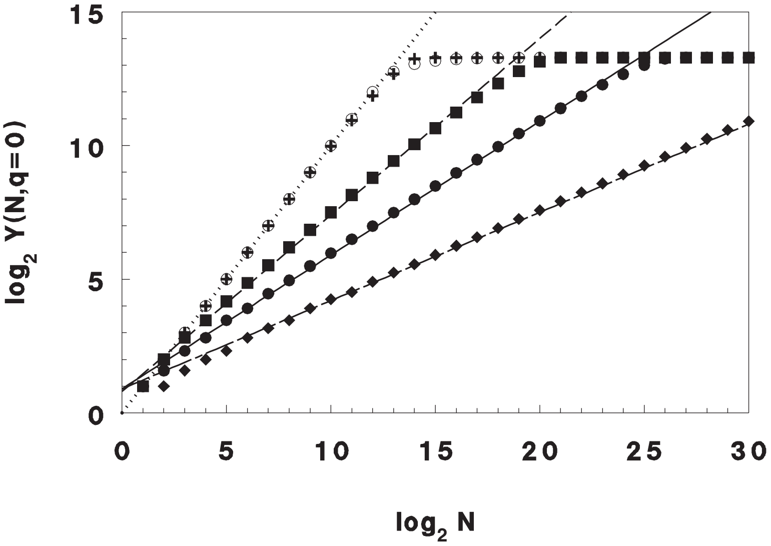

Figure 1. Log–log plots and analytical predictions (lines) for d_0 with $x_n = 1/n$ (full circles and solid line), $1/n^{1/2}$ (squares and dashed line), $1/n^2$ (diamonds and dashed-dotted line), and $n^{1/2}$ and n^2 (crosses and circles with one dotted line for both).

n_{tot}. Equivalently, the measure (2) is not well defined. In standard applications of the BC method one has a fixed number of data points ($n_{\text{tot}} = \text{const}$) and the large N limit is estimated. In this case, for $q \leqslant 0$, one can estimate the sum in (1) as in the derivation of (8), by taking the partial sum with bins containing only one data point. Namely,

$$\frac{1}{1-q} \ln \sum_{i=1}^{N_{\text{sngl}}} p_i^q = \frac{1}{1-q} \ln \left[N^{\frac{1}{1+a}} \left(\frac{1}{n_{\text{tot}}} \right)^q \right]$$
$$= \text{const} + \frac{1}{1-q} \frac{1}{1+a} \ln N.$$

The upper limit can be estimated assuming an equal number of data points in the remaining bins ($N_{\text{r}} \sim N^{1/a}$). One should remember that due to the limited number of data points the number of bins cannot be too large: $1 \ll N < n_{\text{tot}}^{1+a}$. For finer partitions we reach the saturation point; there is a constant number of non-empty bins with exactly one data point inside which corresponds to the value of $\log Y_{\max} = \log n_{\text{tot}}$ (see figures 1 and 2(a), where $\log_2 Y_{\max} = \log_2 10^4 \simeq 13.3$). Finally, we obtain an analytical estimate for large N

$$d_q = \frac{1}{1-q} \frac{1}{1+a} \qquad (q \leqslant 0). \tag{11}$$

For $q > 0$, estimates become more complicated as truncation of the sum can make it smaller than one causing the logarithm to have a change of sign. However, for large q one obtains fast convergence $d_q \to 1$ ($q \to +\infty$). Again, as is clear from figure 2(a), we obtain very good linear fits throughout about ten binary orders of magnitude which is usually interpreted as a sign of fractality and is an excellent agreement with the theoretical estimate.

It has been proven that for fractal set Renyi exponents d_q the HP inequality holds [10]

$$d_q \leqslant d_{q'} \qquad \text{for} \quad q > q'. \tag{12}$$

However, in our case the calculated scaling exponents apparently violate (12) as can be seen from (11) and from figures 2(a) and 3 (full circles and dashed curve).

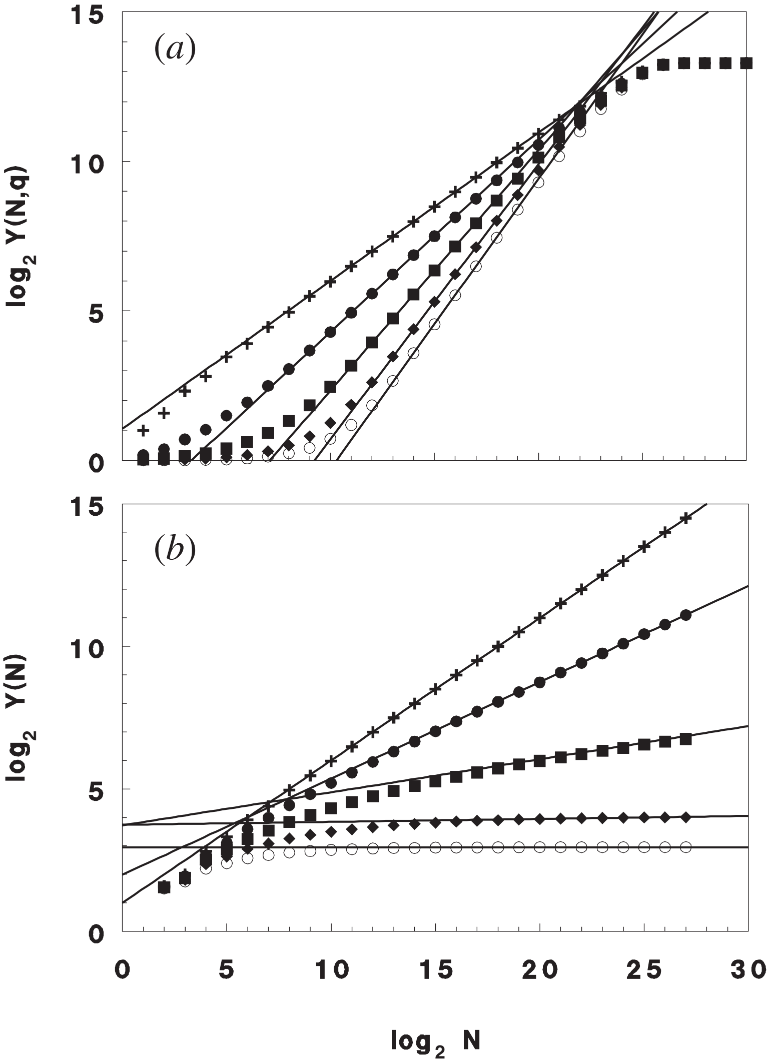

Figure 2. Log–log plots for the harmonic series ($x_n = 1/n$) with $q = 0$ (crosses), 0.25 (full circles), 0.5 (squares), 0.75 (diamonds) and 1 (circles) with corresponding linear fits (solid lines): (*a*) 10^4 data points ($n_{\rm tot}$ = const) and the standard BC method; (*b*) the modified BC algorithm.

We note that when calculating d_q analytically, for well-defined fractal sets such as the triadic Cantor set, the resolution for counting data points increases when the bin number increases. Also, the resolution at a given step (for a given partition) is equal to the bin size (smallest void intervals are of the bin size). This is in contrast to the standard version of the BC method, where the data set is fixed during the whole procedure. Now, we modify the BC method by taking into account for a given partition only those points that are separated from each other by at least the (current) bin size (i.e. the bin size fixes the resolution). This makes the computation more involved and time consuming but, in effect, one can recover the HP inequality.

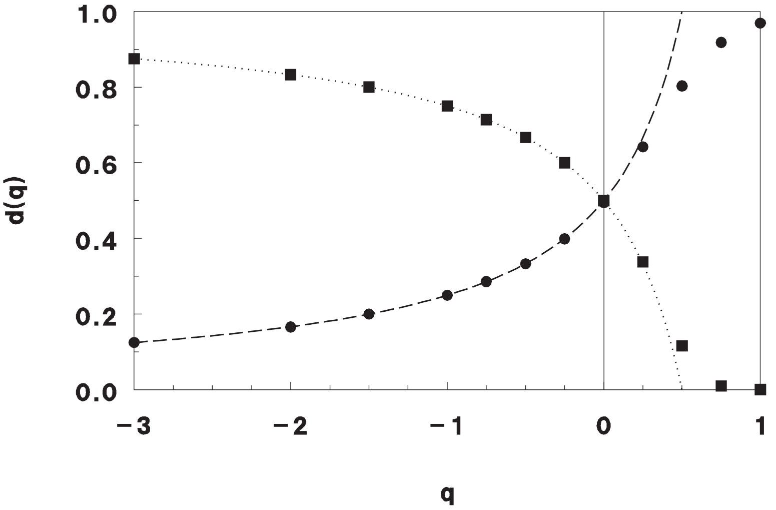

Figure 3. $d(q) \equiv d_q$ computed for the harmonic series ($x_n = 1/n$) with 10^4 data points for the BC (full circles) and modified BC (full squares) methods. The dashed and dotted curves represent analytic estimates (11) and (13), respectively.

For the modified BC method, in a similar way as for the estimate (11), one obtains the following analytical formula:

$$d_q = \frac{1}{1-q}\left[\frac{1}{1+a} - \frac{q}{a}\right] \qquad (q \leqslant 0). \tag{13}$$

In addition, one has $d_q \to 0$ for $q \to +\infty$ (clearly (13) satisfies (12)). This result can be validated numerically, as is displayed in figure 3 (full squares and the dotted curve for equation (13)). Again, we have a very good scaling and linear fit. For positive q this method gives d_q that tends to zero quite fast (while it reaches one slightly more slowly for the standard BC algorithm). Hence, for pseudofractals, the standard and modified BC algorithms give different results due to the ambiguity mentioned earlier. However, in both cases a good scaling and linear fit is obtained.

Our conclusions remain unchanged for sets with an arbitrary number of accumulation points. In particular, the union of two sets with scaling exponents $d_0^{(1)}$, $d_0^{(2)}$ in the large number of bins (N) limit gives

$$\begin{aligned} d_0 &= \frac{1}{\ln N}\, \ln\left[N^{d_0^{(1)}} + N^{d_0^{(1)}}\right] \\ &= \max\left\{d_0^{(1)}, d_0^{(2)}\right\} + \frac{1}{\ln N}\frac{1}{N^{|d_0^{(1)} - d_0^{(2)}|}} \to \max\left\{d_0^{(1)}, d_0^{(2)}\right\}. \end{aligned} \tag{14}$$

As for regular fractals, the scaling exponent of the union is equal to the maximal exponent of the two sets. However, it should be noted that convergence to this result (in the large N limit) is slowest (logarithmic) for $d_0^{(1)} = d_0^{(2)}$. As the number of non-empty bins is greater than for a single set, the points in the log–log plot will be higher for N which is not too large. Hence, the whole plot will be a bit less steep, and for data sets which are not large enough this can lead to lower estimates of the Renyi exponents and a worse linear fit. Details of this effect depend on the particular distribution of both sets in the embedding interval and have been verified numerically. For sets which are not large enough the scaling can be completely lost.

4. Pseudofractals and fat tails

One is often interested in the probability distribution for a large series of data. In particular, in recent years there has been a great interest in the so-called Pareto or fat tails [13], where histograms built out of the data have inverse power-law tails

$$P(x) \sim 1/x^{\beta} \tag{15}$$

where $P(x)$ is the probability distribution. Here we show that non-trivial ($0 < d_0 < 1$) pseudofractals do have this property.

As for the histogram, the time ordering of the data points can be neglected, let us consider for simplicity a monotonic series $\{x_n : x_{n+1} \leqslant x_n\}$. To satisfy (15) the number of data points (Δn) in the interval $[x_{n+\Delta n}, x_n]$ must be

$$\Delta n = \int_{x_{n+\Delta n}}^{x_n} P(x)\,\mathrm{d}x = \frac{C}{\beta}\left[\frac{1}{x_{n+\Delta n}^{\beta-1}} - \frac{1}{x_n^{\beta-1}}\right]$$

where $C > 0$ is a normalization constant. Substituting $C_1 = \beta/C > 0$ and $f(n) \equiv 1/x_n^{\beta-1}$ this yields the following simple linear first-order difference equation:

$$C_1\,\Delta n = f(n+\Delta n) - f(n)$$

with the general solution $f(n) = C_1\,n + C_2$ or, equivalently

$$x_n = \frac{1}{[C_1\,n + C_2]^{\frac{1}{\beta-1}}}.$$

For tails ($n \gg C_2/C_1$ but still far from the accumulation point) the constant C_2 can be neglected and finally we have the asymptotic behaviour

$$x_n \sim \frac{1}{n^{\frac{1}{\beta-1}}} \equiv \frac{1}{n^a}. \tag{16}$$

Hence, the tail exponent β can be expressed in terms of a or d_0 as

$$\beta = \frac{a+1}{a} = \frac{1}{1-d_0}. \tag{17}$$

The above formula displays the simple relation between the pseudofractal parameter a, the tail index β and the BC exponent d_0.

5. Conclusions

In this paper, we have investigated general sets with accumulation points, that are not fractals, though they display fractal-like scaling behaviour. The scaling exponent d_0 (equation (3)) as obtained by the BC method is given by equation (10). Furthermore, we have found the analytical formula for d_q (for $q \leqslant 0$) for the inverse power series as given by the standard BC algorithm (equation (11)), that perfectly fits the numerical results (figure 2(*a*)). The obtained exponents violate the HP inequality, which can be viewed as an indicator of the pseudofractal behaviour.

Similar results are obtained for the modified BC algorithm (where the number of data points taken into account increases with the increased resolution), but in this case the HP inequality is preserved (see equation (13) and figure 2(*b*)). Hence, the two schemes give different d_q for the pseudofractal sets. Our results remain valid for sets with an arbitrary number of accumulation points, where the overall scaling exponent is equal to the maximal exponent of constituent sets. Also, in this case, one can observe a worsening of the linear fit.

In general, from the point of view of the fractal properties and the BC methods, there are four types of sets:

(i) *mathematical fractals*—sets that are well defined and their fractal properties can be rigorously proven (i.e. without numerical approximations), such as the triadic Cantor set;

(ii) *physical fractals*—finite sets that are (computer) representations of mathematical fractals. In this case, one obtains good scaling and linear fit with the BC method, HP inequality holds, and both the BC and modified BC methods (described in section 3) give the same results;

(iii) *pseudofractals*—finite sets that are not finite representations of mathematical fractals, though they show good scaling and linear fit with the BC method. The resulting exponents violate the HP inequality and the BC and modified BC algorithms give different values for d_q. The general formula for d_0, when x_n asymptotic is known, is given by equation (10);

(iv) *non-fractals*—i.e. sets for which the BC algorithm does not exhibit any scaling.

The sets of types (i) and (iv) can be easily distinguished. However, it is quite non-trivial to distinguish between sets of types (ii) and (iii). Here, one cannot apply the rigorous mathematical machinery as the whole set is usually unknown. In these cases, numerical methods lead to nice scaling making them impossible to tell apart. In this context, the violation of the HP inequality appears to be a simple and useful indicator, in addition to different results obtained by the standard and modified BC algorithms.

Different classes of non-trivial ($0 < d_0 < 1$) pseudofractals have scaling properties equivalent to the series $\{x_n = 1/n^a\}$. In particular, for $q > 0$, the BC method gives d_q close to the embedding dimension while for the modified BC algorithm d_q approaches zero. For $q \leqslant 0$, analytical formulae for d_q are given by equations (11) and (13), respectively.

Finally, as shown by equation (17), the parameter a of the pseudofractal series is simply related to the tail index β, as well as to the BC Renyi exponent d_0. This means that histograms made of non-trivial pseudofractal sets have Pareto (fat) tails. This relation is another signal of possible pseudofractality.

References

[1] Mandelbrot B B 1977 *Fractals: Form, Chance, and Dimension* (San Francisco: Freeman)

[2] Białas A and Peszanski R 1988 *Nucl. Phys.* B **308** 803

[3] Jones B J T, Martinez V T, Saar E and Einasto J 1988 *Astrophys. J.* **332** L1
Coleman P H, Pietronero L and Sanders R H 1988 *Astron. Astrophys.* **200** L32

[4] Hofstadter D R 1976 *Phys. Rev.* B **14** 2239

[5] Evertsz C J G 1995 Self-similarity of high-frequency USD–DEM exchange rates *Proc. 1st Int. Conf. on High Frequency Data in Finance (Zürich)*

[6] Molteno T C A 1993 *Phys. Rev.* E **48** R3263

[7] Farmer J D, Ott E and Yorke J A 1983 *Physica* D **7** 153

[8] Kolmogorov A N 1958 *Dokl. Akad. Nauk SSSR* **119** 861

[9] Renyi A 1970 *Probability Theory* (Amsterdam: North-Holland)

[10] Hentschel H G E and Procaccia I 1983 *Physica* D **8** 435

[11] Badii R and Politi A *J. Stat. Phys.* **40** 725 appendix A

[12] Górski A Z 1998 Comment on fractality of quantum mechanical energy spectra *Preprint* arXiv chao-dyn/9804034

[13] Mantegna R N and Stanley H E 2000 *Econophysics: Correlations and Complexity in Finance* (Cambridge: Cambridge University Press)
Mantegna R N and Stanley H E 1995 *Nature* **376** 46