FINITE SAMPLE CORRECTIONS TO ENTROPY AND DIMENSION ESTIMATES

Peter GRASSBERGER
Physics Department, University of Wuppertal, D-5600 Wuppertal 1, FRG

Received 17 December 1987; revised manuscript received 9 February 1988; accepted for publication 10 February 1988
Communicated by A.P. Fordy

We derive the systematic corrections to estimates of generalized (Renyi) entropies and to generalized dimensions D_q from finite data sets. As an application, we discuss correlation estimates of D_q for the Hénon map. We end with some remarks about lacunarity measures.

The last years have witnessed an increasing number of attempts to estimate fractal dimensions and entropies of dynamical systems, both of models and of real systems. In addition to the fractal dimension proper and to the metric entropy, much attention has also been devoted to Renyi-type [1] entropies and to the Renyi dimensions associated to them ("generalized dimensions" [2]), and to their Legendre transforms ("$f(\alpha)$-spectra" [3]).

In most of these analyses, the "experimental" data consisted of a sample of points which are supposed to be randomly distributed according to some measure μ. The dimensions and entropies are defined with respect to μ, and it is obvious that the statistical fluctuations of the sample will induce both statistical and systematic deviations of the estimated values. It is the purpose of this note to discuss the latter (i.e. the systematic) errors. As we shall see, for all presently employed methods the dominant corrections are fairly simple and universal. They are straightforward generalizations of and improvements over the correction derived recently by Herzel [4] for Shannon entropies.

Let us consider a measure μ and a partition C of its support such that the ith element of the partition carries a weight p_i. The order-q entropies of $\{\mu, \mathrm{C}\}$ are defined as [1]

$$h_q(\mu, \mathrm{C}) = \frac{1}{1-q} \log \sum_i p_i^q \,. \tag{1}$$

The partition here is completely arbitrary. In order to get the dynamic entropies of a dynamical system, we first introduce a symbolic dynamics, i.e. we represent each trajectory by a symbol sequence. We then take partitions C_T such that each set corresponds to a different string of symbols in some fixed time window $[t, t+T]$ of length T. The order-q dynamic entropy is finally defined as

$$h_q(\mu) = \lim_{T\to\infty} \{h_q(\mu, \mathrm{C}_{T+1}) - h_q(\mu, \mathrm{C}_T)\} \,. \tag{2}$$

If we have a family of partitions $\mathrm{C}(\epsilon)$ with the property that each element of $\mathrm{C}(\epsilon)$ has a characteristic length ϵ (we assume now that the support of μ is embedded in some euclidean space), we can define order-q dimensions by [2]

$$D_q = \lim_{\epsilon\to 0} \frac{-h_q(\mu, \mathrm{C}(\epsilon))}{\log \epsilon} \,. \tag{3}$$

A more general definition of D_q was proposed in ref. [5] (and used later in ref. [6]): instead of partitions with fixed characteristic length ϵ, we follow the definition of Hausdorff–Besicovitch [7] in using partitions whose elements have diameters $\delta_i \leqslant \epsilon$, and define D_q such that

$$\sum_i \frac{p_i^q}{\delta_i^\tau} \underset{\epsilon\to 0}{\sim} 1 \tag{4}$$

for $\tau = (q-1)D_q$. For a more rigourous definition, one would have to take suprema respectively, infima with respect to all partitions with $\delta_i \leqslant \epsilon$ [5], but we shall not go into detail. Instead, we point out that eq.

0375-9601/88/$ 03.50

(4) goes over into eq. (3) when $\delta_i = \epsilon$ for all i.

Eq. (4) can be generalized to the case where the sum does not extend over the sets of a partition which covers the support of μ completely and uniquely. In particular, we can instead take balls of radius δ_i centered at M μ-randomly chosen points x_i. In this case, the sum in eq. (4) has to be modified to

$$\frac{1}{M}\sum_{i=1}^{M}\frac{p_i^{q-1}}{\delta_i^{\tau}} \underset{\epsilon\to 0}{\sim} 1 . \tag{5}$$

The above considerations have suggested (at least) 3 different methods of estimating $D_q = \tau(q)/(q-1)$ (and, via a Legendre transform $f(\alpha) = q\,\mathrm{d}\tau/\mathrm{d}q - \tau(q)$, the Hausdorff dimensions $f(\alpha)$ of the sets of points with pointwise dimension α [3,6]):

(α) box counting counting: take a fixed mesh with box size ϵ, and apply eq. (3). The probabilities p_i are estimated from N random points as

$$p_i \approx n_i/N , \tag{6}$$

where n_i is the number of points in box i.

(β) correlation sums: instead of taking a fixed mesh, count the numbers n_i of points (out of N in total) falling in balls of radius ϵ around M randomly chosen centers and use eq. (5), estimating again p_i from eq. (6) [2]. In the simplest case, one takes $M = N$ and uses the same set of points as centers and as correlates [8–10]. Notice that the point defining the center of the ball is *not* included in n_i.

(γ) nearest neighbour distances: after having put M random centers and N correlate points, one computes the distance $r_i^{(j)}$ of the jth nearest neighbours of the center x_i, among all correlate points. From this, one estimates the size δ_i of a ball with measure $p = j/N$ around x_i as $\delta_i \approx r_i^{(j)}$, and uses then eq. (5) [11,5].

For method (γ), the systematic corrections due to the finiteness of N are well known [5,12]. When $j \ll N$, one has

$$\overline{(r^{(j)})^{-\tau}} \equiv \frac{1}{M}\sum_{i=1}^{M}(r_i^{(j)})^{-\tau} \sim N^{q-1}\frac{\Gamma(j+1-q)}{\Gamma(j)} \tag{7a}$$

and, from the limit $\tau \to 0$,

$$D_1\,\overline{\log r^{(j)}} = \Psi(j) - \log N \tag{7b}$$

($\psi(x) = \mathrm{d}\log\Gamma(x)/\mathrm{d}x$). For $j \gg 1$ this goes over to eq. (5), but it can deviate considerably for small j.

Methods (α) and (β) as well as the entropy estimate in eq. (1) require that powers or logarithms of probabilities p_i are estimated from the numbers n_i of points falling into domain i, out of a sample of N points in total. We shall discuss here only the case $p_i \ll 1$, i.e. of very small domains. In successive realizations, the numbers n_i of points falling into the domain would be distributed according to a Poisson distribution with $\langle n_i \rangle = Np_i$. Here and in the following, we use brackets $\langle\ \rangle$ to indicate averages over different realizations, keeping i fixed. This is to be distiguished from averages over i which are indicated by overbars in eq. (7).

The systematic corrections in methods (α) and (β) are then immediately obtained for integer positive q, using the well known formula (here and in some subsequent equations we omit the index i)

$$\langle n \rangle^{q} = \left\langle \frac{n!}{(n-q)!} \right\rangle, \quad q \geqslant 1 . \tag{8}$$

The logarithm in eq. (1) becomes then e.g.

$$\log\left\langle N^{-q}\sum_i \frac{n_i!}{(n_i-q)!}\right\rangle .$$

Since the quantity in brackets is now itself an average over many essentially independent terms, its fluctuactions are strongly suppressed. We can thus drop the brackets, and the correct estimate based on eqs. (1) and (8) becomes

$$h_q(\mu, \mathrm{C}) \approx \frac{1}{1-q}\log N^{-q}\sum_i \frac{n_i!}{(n_i-q)!}$$

(q integer, $\geqslant 2$). (9)

For non-integer q, things are not so simple. One might guess that the generalization of eq. (8) is simply obtained by replacing the factorials by Γ-functions, but this is not true. Indeed, we cannot expect to find coefficients $\Phi_n(q)$ such that for a Poisson distribution

$$\langle n \rangle^{q} = \langle \Phi_n(q) \rangle \equiv \mathrm{e}^{-\langle n \rangle}\sum_{n=0}^{\infty}\Phi_n(q)\frac{\langle n \rangle^{n}}{n!} \tag{10}$$

exactly and absolutely convergent for all real q, since the sum on the r.h.s. would otherwise be the (nonexisting) Taylor sum of $z^q \mathrm{e}^z$ with $z = \langle n \rangle$ [#1]. But we

[#1] This simple argument was provided by K. Eilenberger (private communication).

can find $\Phi_n(q)$ such that eq. (10) holds asymptotically for large n, and is a good approximation for finite n.

Indeed, using eqs. 6.1(1), 6.7(7), 6.5(3), and 6.13.1(1) of ref. [13], we can express $\langle n\rangle^q \mathrm{e}^{\langle n\rangle}$ as the expectation value of $\Gamma(n+1)/\Gamma(n-q+1)$ plus terms $\sim\langle n\rangle^{-k}$, $k=1,2,\ldots$. The latter can be transformed into expectation values of $(n+l)^{-1}$, $l=1,2,\ldots$, plus terms exponentially small in $\langle n\rangle$, and finally we obtain

$$\langle n\rangle^q = \left\langle \frac{\Gamma(n+1)}{\Gamma(n-q+1)} + \frac{(-1)^n}{\pi}\sum_{k=1}^{R}\frac{\Gamma(k+q)\sin[\pi(k+q)]}{(n+k)\Gamma(k)}\right\rangle + o(\langle n\rangle^{-R}\mathrm{e}^{-\langle n\rangle}) \tag{11}$$

for any integer R. We see that indeed for large R the contributions of each fixed n alternate in sign, introducing thereby large statistical errors and rendering the expression with very large R useless for numerical purposes. We found that the first nontrivial order,

$$\langle n\rangle^q \approx \left\langle \frac{\Gamma(n+1)}{\Gamma(n-q+1)} - \frac{(-1)^n\Gamma(1+q)\sin\pi q}{\pi(n+1)}\right\rangle, \tag{12}$$

gave numerically the most reasonable results. The systematic error committed here is smaller than without the second term both for large and small $\langle n\rangle$, contributions with different n do not yet cancel strongly, and numerical simulations show that eq. (12) gives relative errors $\lesssim 10^{-3}$ for all $q\geqslant 0$, down to $\langle n\rangle\approx 3$.

With eqs. (1) and (12), the estimate of the Shannon entropy $h=\lim_{q\to 1}h_q$ becomes

$$h(\mu,\mathrm{C}) \approx \sum_i \frac{n_i}{N}\left(\log N - \Psi(n_i) - \frac{(-1)^{n_i}}{n_i+1}\right). \tag{13}$$

Using the estimate $\Psi(x)\approx\log x - 1/2x$ for large x (ref. [13], eq. 1.18(7)), we see that we not only obtain the usual expression when all $n_i\gg 1$, but that the first correction is indeed as derived by Herzel [4]. In order to show the improvement due to the higher order correction terms, we present in fig. 1 the errors in estimating the entropy of a binary random sequence (1 bit/digit) using eq. (2), and improving

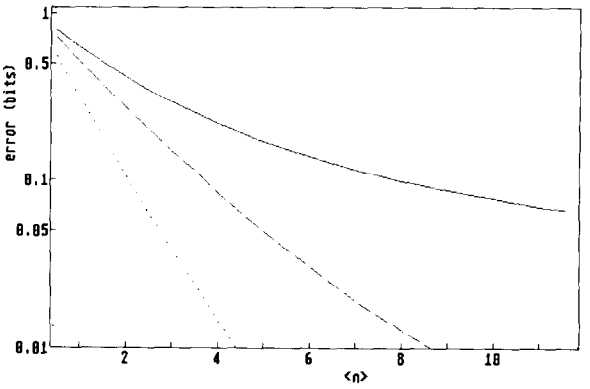

Fig. 1. Errors in estimating the entropy of a random binary string as $h\approx h_{T+1}-h_T$, versus the average number $\langle n\rangle$ of strings with definite symbol sequence of length T. Full line: p_i estimated as n_i/N; dashed line: correction of ref. [4] included; dotted line: eq. (13).

successively over the naive estimate $p_i\approx n_i/N$. We see that indeed the error is smallest with eq. (13), and is negligible for most purposes when $\langle n\rangle\geqslant 3$.

In fig. 1 we see also that all approaches yield wrong estimates when $\langle n\rangle\to 0$. This has to be so, as in this limit no entropy estimate can be possible from eq. (2). It again demonstrates what we have said above, that there cannot be an exact representation as in eq. (10).

The analogous expressions following from eqs. (3) and (4) are straightforward. For dimension estimates from correlation sums (method (β)) we obtain

$$\tau(q) \approx (\log\epsilon)^{-1}\log\left[\frac{N}{MN^q}\sum_{i=1}^{M}\left(\frac{\Gamma(n_i+1)}{\Gamma(n_i+2-q)} + \frac{(-1)^{n_i}}{(n_i+1)\Gamma(1-q)}\right)\right]. \tag{14}$$

This simplifies most for $q=2$, giving there the usual expression for the correlation dimension [8–10]. That the case $q=2$ is peculiar in showing no finite sample corrections was pointed out in ref. [5].

For the information dimension ($q\to 1$), the correlation method gives in particular

$$D_{\mathrm{info}} \approx (M\log\epsilon)^{-1} \times \sum_{i=1}^{M}\left(\Psi(n_i+1) - \log N - \frac{(-1)^{n_i}}{n_i+1}\right). \tag{15}$$

Using the correlation method for D_{info} or for D_q

with $q<1$ naively, one encounters problems as soon as one finds a center point in whose neighbourhood there are no other points ($n_i=0$). This has led some authors [14] to replace n_i by n_i+1, i.e. to include the center point when counting the number of points in a small ball. Following them, other authors [15] have included the center point even in estimates of D_q with $q\geqslant 2$. This is by no means justified, and it has led to wrong conclusions. Since $\Psi(x)$ is finite at $x=1$, all summands in eq. (15) are finite and we do not encounter this problem for D_{info}. Indeed, we have no similar problem for any of the D_q with $q<1$ either.

The corrections discussed above should be most important for multifractals with a wide $f(\alpha)$-spectrum since there it is virtually impossible to have large numbers of points in all elements of a fine partition. In order to test this, we performed dimension estimates for the Hénon map [16] with the usual parameters $a=1.4$, $b=0.3$. In fig. 2 we show effective dimensions estimated by the correlation method,

$$D_q^{\text{eff}}(\epsilon)=(q-1)^{-1}\log_4\left\{\sum_{i=1}^{M}\left(\frac{\Gamma(\tilde{n}_i+1)}{\Gamma(\tilde{n}_i+2-q)}+\frac{(-1)^{\tilde{n}_i}}{(\tilde{n}_i+1)\Gamma(1-q)}\right)\left[\sum_{i=1}^{M}\left(\frac{\Gamma(n_i+1)}{\Gamma(n_i+2-q)}+\frac{(-1)^{n_i}}{(n_i+1)\Gamma(1-q)}\right)\right]^{-1}\right\}, \tag{16}$$

where $\tilde{n}_i$ (respectively n_i) is the number of points y_k ($k\leqslant N$) less that 2ϵ (respectively $\epsilon/2$) away from x_i. As metric we used the maximum norm, and we employed a modified box-assisted algorithm [17] to achieve very high statistics ($M=2.2\times10^5$, and $N=4\times10^6$ at the smallest scale) on an Atari ST home computer. There are still further finite-N corrections for $q=0$ and $q=1$, while there are none for $q\geqslant 2$. But they appear very late indeed. We have indicated by an arrow the value of ϵ below which we had central points without any neighbours closer than $\epsilon/2$. The dimension estimates are still correct substantially below these values also for $q=0$ and $q=1$, while they would have gone astray there, at latest, if we had not included the corrections.

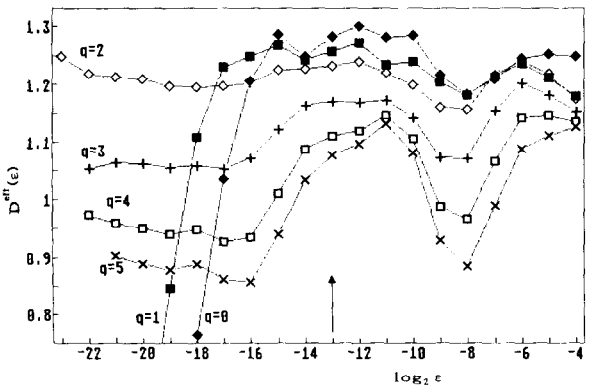

Fig. 2. Dimension estimates (using eq. (16)) for the Hénon map for $q=0$, 1, 2, 3, 4, and 5 versus resolution ϵ, using $M=2.2\times10^5$ central points and 4×10^6 correlate points. The arrow indicates that ϵ-value for which no correlates were found for some central points at distances $\leqslant\epsilon/2$.

We should finally point out the fluctuations in $D^{\text{eff}}(\epsilon)$, first observed in ref. [18] (see also refs. [19,20]). They are not statistical but are lacunarity effects as explained in refs. [10,21] in a similar case. It is however not possible to use them as a measure of lacunarity (as proposed in ref. [22]) since they are smeared out for $\epsilon\to0$ since the Hénon attractor is not *strictly* self-similar (for some fractals such as e.g. DLA clusters [8], no fluctuations are observed at all, although they are obviously lacunar). This attenuation of fluctuations in the scaling limit is clearly seen in fig. 2. The amplitude of the fluctuations increases strongly with q [20], since for large q only few regions of the attractor contribute substantially. For $0\lesssim q\lesssim 2$, regions with different local lacunarity smear out the fluctuations already at rather large ϵ. A lacunarity measure not suffering from this problem was proposed in ref. [23] (notice that this definition was later "corrected" by the same authors in ref. [24] in a wrong way [25]).

Summarizing, we have derived the systematic corrections in estimates of (Renyi and Shannon) entropies, and thus also of generalized fractal dimensions. They can be quite important, but fortunately the leading terms can be taken into account systematically with very minor efforts. Together with the trick of ref. [17], they allowed us to reject a recent conjecture how to measure lacunarity.

I am very much indebted to Dr. Herzel for sending me his preprint wich indeed stimulated the present investigation, and to Professors K. Eilenberger and F. Krause for most useful discussions.

References

[1] A. Renyi, Probability theory (North-Holland, Amsterdam, 1971).

[2] B.B. Mandelbrot, The fractal geometry of nature (Freeman, San Francisco, 1982);
P. Grassberger, Phys. Lett. A 97 (1983) 227.

[3] R. Benzi, G. Paladin, G. Parisi and A. Vulpiani, J. Phys. A 17 (1984) 3521;
E.B. Vul, Ya. G. Sinai and K.M. Khanin, Russ. Math. Surv. 39 (1984) 1;
B.B. Mandelbrot, in: Lecture notes in Mathematics, Vol. 565. Turbulence and the Navier–Stokes equations, ed. R. Temam (Springer, Berlin, 1975).

[4] H. Herzel, Complexity of symbol sequences, Berlin preprint (1987).

[5] P. Grassberger, Phys. Lett. A 107 (1985) 101.

[6] T. Halsey et al., Phys. Rev. A 33 (1986) 1141.

[7] K.J. Falconer, The geometry of fractal sets (Cambridge Univ. Press, Cambridge, 1985).

[8] T. Witten and L. Sander, Phys. Rev. Lett. 47 (1981) 1400.

[9] F. Takens, Invariants related to dimension and entropy, in: Atas do 13° Coloquio Brasileiro de Matematica (1983).

[10] P. Grassberger and I. Procaccia, Physica D 9 (1983) 189.

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

[12] R.L. Somorjai, Methods for estimating the intrinsic dimensionality of high-dimensional point sets, in: Dimensions and entropies in chaotic systems, ed. G. Mayer-Kress (Springer, Berlin, 1986).

[13] A. Erdelyi et al., Higher transcendental functions, Vol. 1 (McGraw-Hill, New York, 1953).

[14] A. Cohen and I. Procaccia, Phys. Rev. A 31 (1985) 1872.

[15] J. Holzfuss and G. Mayer-Kress, An approach to error estimation in the application of dimension algorithms, in: Dimensions and entropies in dynamic systems, ed. G. Mayer-Kress (Springer, Berlin, 1986);
K. Fraedrich, J. Atmos. Sci. 43 (1986) 419.

[16] M. Hénon, Commun. Math. Phys. 50 (1976) 69.

[17] J. Theiler, Phys. Rev. A 36 (1987) 4456.

[18] P. Grassberger, Phys. Lett. A 97 (1983) 224.

[19] W.E. Caswell and J.A. Yorke, Invisible errors in dimension calculations, in: Dimensions and entropies in dynamic systems, ed. G. Mayer-Kress, (Springer, Berlin, 1986).

[20] A. Arneodo, G. Grasseau and E.J. Kostelich, preprint (1987).

[21] R. Badii and A. Politi, Phys. Lett. A 104 (1984) 303.

[22] L.A. Smith, J.D. Fournier and E.A. Spiegel, Phys. Lett. A 114 (1986) 465.

[23] Y. Gefen, A. Aharony and B.B. Mandelbrot, Phys. Rev. Lett. 50 (1983) 145.

[24] Y. Gefen, A. Aharony and B.B. Mandelbrot, J. Phys. A 17 (1984) 1277.

[25] B. Lin and Z.R. Yang, J. Phys. A 19 (1986) L49.