JOURNAL OF COMPUTATIONAL PHYSICS 99, 196–202 (1992)

Box-Counting Algorithm and Dimensional Analysis of a Pulsar

F. H. LING* AND G. SCHMIDT

Department of Physics and Engineering Physics, Stevens Institute of Technology, Hoboken, New Jersey 07030

Received March 3, 1990; revised March 4, 1991

It is argued that the use of the box-counting algorithm to calculate the correlation dimension is a better choice than the Grassberger–Procaccia (correlation integral) algorithm for dealing with an experimental data set. This is illuminated by treating three classical examples: the logistic map, the Hénon map and the Lorenz equation. The intensity data of a pulsar is also treated which is revealed to have a least embedding dimension of 14 and the correlation dimension of about 4.5. © 1992 Academic Press, Inc.

Chaotic systems are typically studied as follows. First, a time series is generated by measuring a chaotic quantity at times $t = 0, \tau, 2\tau, ..., n\tau, ...$. Occasionally, nature provides the time series directly as in the case of pulsar data to be investigated later in this paper.

Second, one chooses an embedding dimension d, and constructs vectors $\mathbf{X}_1, ..., \mathbf{X}_N$ in this d-dimensional space from the time series data using a phase recovering technique [1–4].

Finally, if the set of points $\mathbf{X}_i$ belongs to a strange attractor, one may find the generalized dimensions

$$D_q = \frac{1}{1-q} \lim_{\varepsilon \to 0} \frac{\ln \sum_{i=1}^{N(\varepsilon)} P_i^q(\varepsilon)}{\ln(1/\varepsilon)}, \quad q \geqslant 0, q \neq 1, \tag{1}$$

where $N(\varepsilon)$ is the number of non-empty boxes, $P_i(\varepsilon)$ is the probability of finding a point in a d-dimensional box with length scale ε, and the summation is carried out over all boxes. However, the limit value is in general not available for the real system, so one plots the numerator of Eq. (1) versus the denominator for finite ε. If a reasonably good straight line fits the discrete points in a certain range of ε, the slope of this straight line determines D_q. Since the least embedding dimension is in general unknown, one experiments by raising the value of d, looking for saturation of D_q.

The most frequently used dimensions are the capacity dimension

$$D = D_0 = \lim_{\varepsilon \to 0} \frac{\ln N(\varepsilon)}{\ln(1/\varepsilon)}, \tag{2}$$

the information dimension

$$\nu = \lim_{q \to 1} D_q = \lim_{\varepsilon \to 0} \frac{\sum_{i=1}^{N(\varepsilon)} P_i(\varepsilon) \ln P_i(\varepsilon)}{\ln(1/\varepsilon)}, \tag{3}$$

and the correlation dimension

$$\nu = D_2 = \lim_{\varepsilon \to 0} \frac{\ln \sum_{i=1}^{N(\varepsilon)} P_i^2(\varepsilon)}{\ln \varepsilon}. \tag{4}$$

Grassberger and Procaccia have introduced the correlation integral [5–8]

$$C(\varepsilon) = \lim_{M \to \infty} \frac{\sum_{i,j=1, i \neq j}^{M} H(\varepsilon - \|\mathbf{X}_i - \mathbf{X}_j\|)}{M^2}, \tag{5}$$

where $H(\cdot)$ is the Heaviside function and M is the number of points considered. It can be shown that if the set is self-similar then $C(\varepsilon)$ is proportional to $\sum_{i=1}^{N(\varepsilon)} P_i^2(\varepsilon)$, so ν can be evaluated with the help of the correlation integral,

$$\nu = \lim_{\varepsilon \to 0} \frac{\ln C(\varepsilon)}{\ln \varepsilon}. \tag{6}$$

Recently, most authors use this Grassberger–Procaccia algorithm to calculate the correlation dimension as the main measure of a strange attractor, because it is argued that the box-counting algorithm cannot be used to calculate the capacity dimension of a small data set, and it seems that there is also no other suitable algorithm for calculating generalized dimensions. However, the Grassberger–Procaccia algorithm is rather time-consuming, requiring a

* On leave from the Department of Engineering Mechanics, Shanghai Jiao Tong University, Shanghai 200 030, PR China. Present address: Department of Chemistry and Chemical Engineering, Stevens Institut of Technology, Hoboken, New Jersey 07030.

0021-9991/92 $3.00

calculation amount roughly proportional to M^2, so many authors use a modified version; they choose a small number of reference points and only consider the distance from them to the remaining points. It is evident that this variant will bring additional errors in the results as pointed out in [9]. Theiler [10] suggests an algorithm in which a presorting is performed, so that one does not need to consider those points which are far away from the reference points and therefore the computational work is somehow reduced. This algorithm brings additional complexities and is not of obvious advantage except the embedding dimension is very low.

Let us consider the box-counting algorithm more closely. In order to occupy all non-empty boxes with small length scale ε, one needs a huge number of points, or the calculated capacity dimension will be very inaccurate. Noting this point, a very early paper [11] already shows that box-counting algorithms are generally impractical for calculating the capacity. However, if one uses box-counting algorithms to determine the correlation dimension, then the crucial thing is to calculate the value of the sum of the squared probability $\sum_{i=1}^{N(\varepsilon)} P_i^2(\varepsilon)$, which is far less sensitive to the non-sufficient occupation with respect to the total number of non-empty boxes $N(\varepsilon)$. Therefore, there is a good reason to expect that one can use a box-counting algorithm to calculate correlation dimension effectively.

Another point which is often mentioned in the literature as a shortcoming of the box-counting algorithm is the requirement of a huge memory for storing all possible boxes in a probably high-dimensional embedding space. But this difficulty can be easily overcome if one only considers occupied boxes, since then the number of storage places will not exceed the number of points in the data set, although a larger calculation amount will be required in return.

An obvious advantage of using the box-counting algorithm to calculate the correlation dimension is that it only requires a calculation amount roughly proportional to $MN(\varepsilon)$, since we only need to check whether a new point is falling into an already occupied box or it creates a new occupied box itself. Typically, the value of $N(\varepsilon)$ is only 10 to 20% of the value of M. Moreover, the computer mainly does comparison work, which is less time consuming than calculating distances in the Grassberger–Procaccia algorithm. Moreover, some tricks help one to raise the efficiency of the box-counting algorithm. For example, the data can be enlarged and rounded to an integer form, which keeps sufficient significant digitals and eases further manipulations. Furthermore, the coordinates of the occupied boxes can be stored in the ASCII code form rather than in the digital form. These measures reduce both the storage requirement and the computer time. Our experience shows that the box-counting algorithm enables one to calculate generalized dimensions of an experimental data set with a personal computer, or with a larger computer, one can do such a calculation on-line, which may be of a practical meaning.

To test the accuracy of the box-counting algorithm

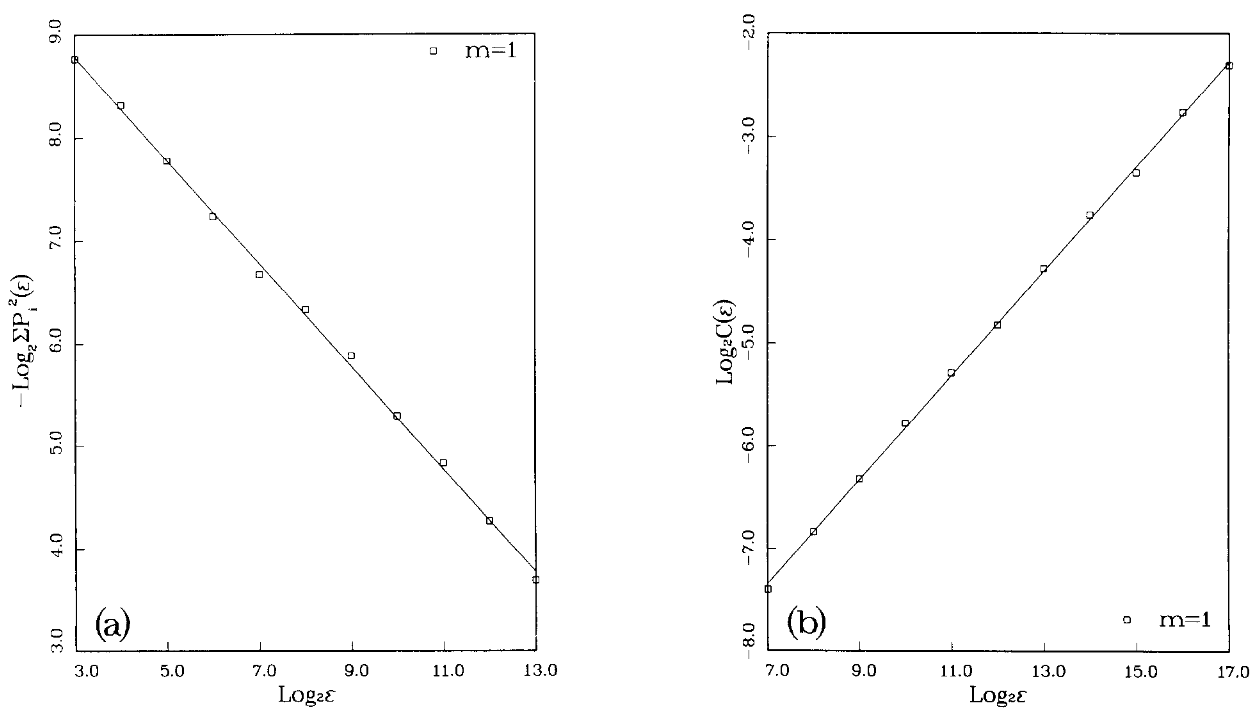

FIG. 1. Correlation dimension of the logistic map with $\lambda = 3.5699456\cdots$: (a) box-counting; (b) Grassberger–Procaccia algorithm.

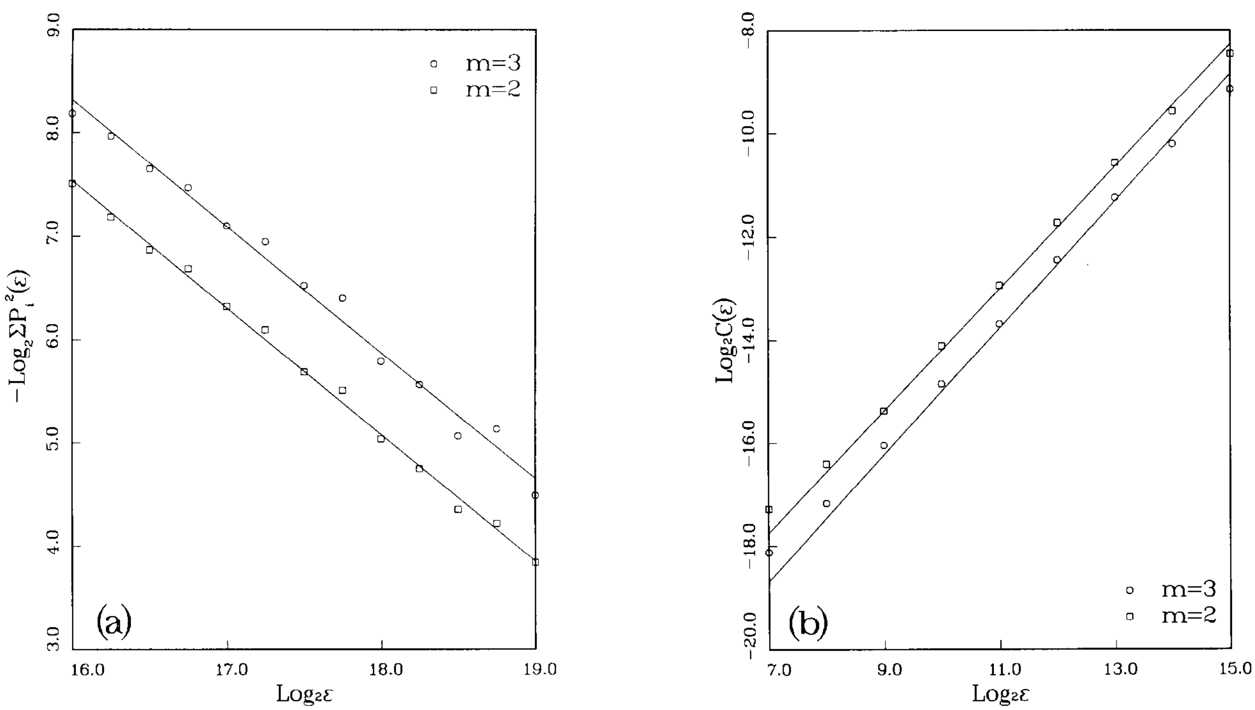

FIG. 2. Correlation dimension of the Hénon map with $a = 1.4$, $b = 0.3$: (a) box-counting; (b) Grassberger–Procaccia algorithm.

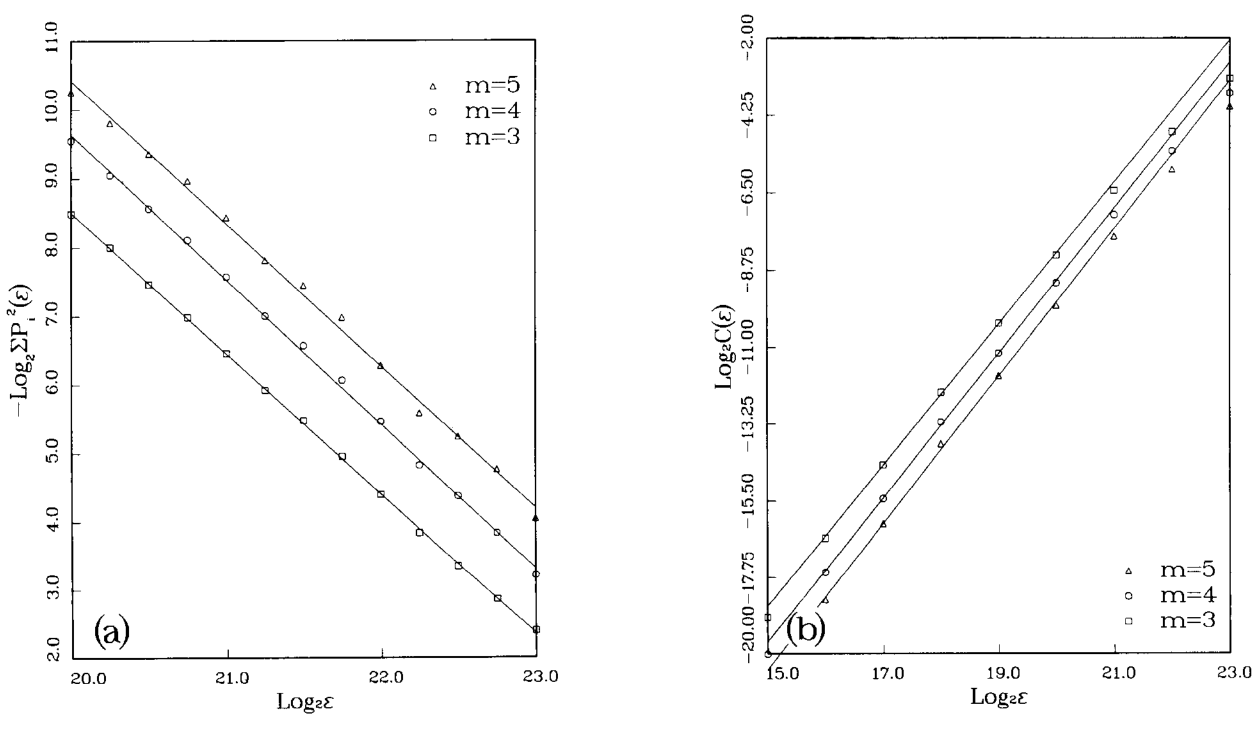

FIG. 3. Correlation dimension of the Lorenz equation with $R = 28$, $\sigma = 10$, $b = \frac{8}{3}$: (a) box-counting; (b) Grassberger–Procaccia algorithm.

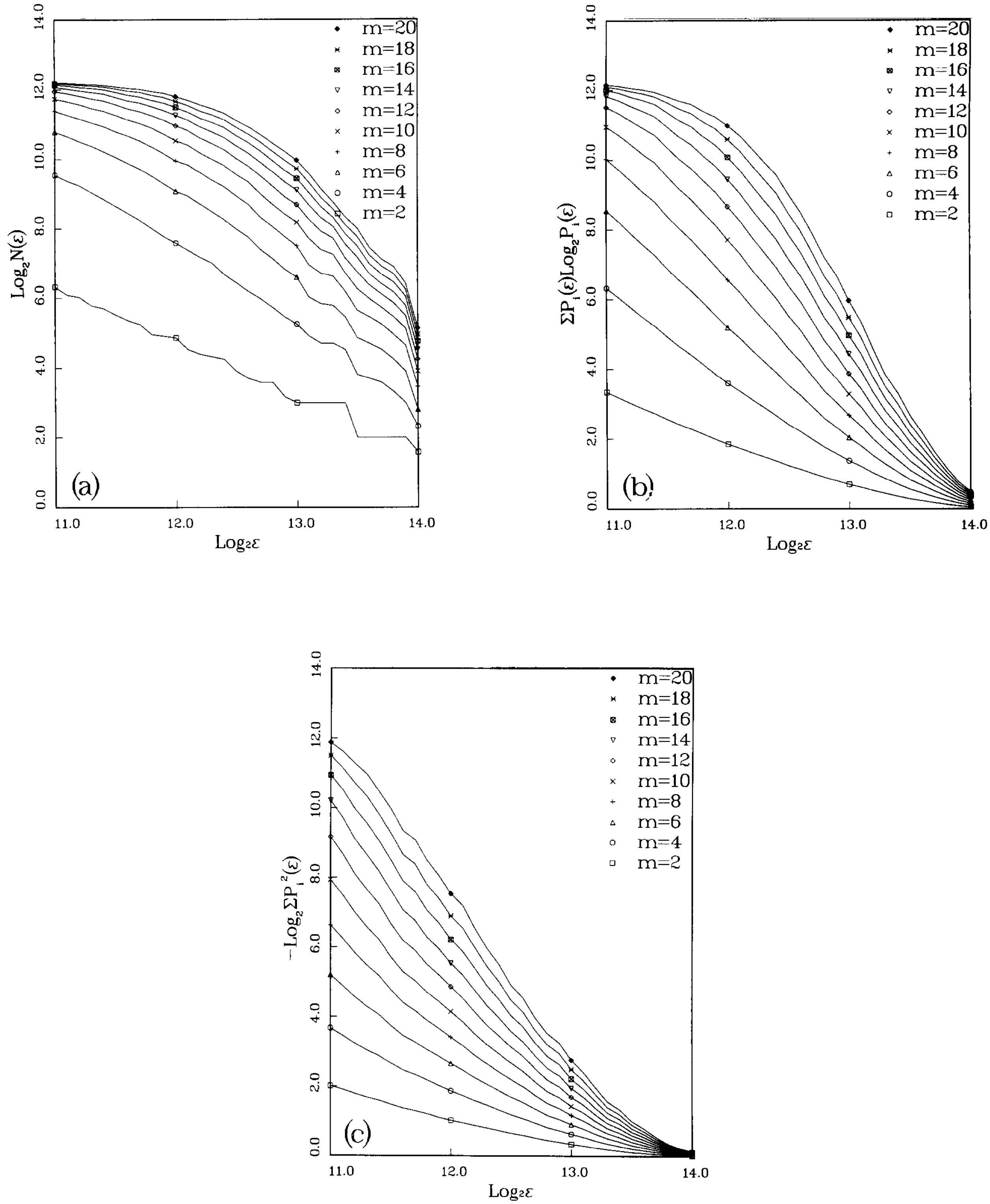

FIG. 4. Generalized dimensions of pulsar 0950 + 08 calculated with box-counting by using 4000 entries: (a) capacity; (b) information dimension; (c) correlation dimension.

TABLE I

Fractal Dimensions Calculated with Box-Counting and the Grassberger–Procaccia Algorithm

	No.of iter.	Time increm.	Emb. dim.	Box-counting	Grassberger–Procaccia	Reference values [7]
Logistic map $\lambda = 3.56699456$	1000		1	$D = 0.538$ $\sigma = 0.518$ $\nu = 0.500$	$\nu = 0.501$	$D = 0.528$ $\sigma = 0.517$ $\nu = 0.500$
Hénon map $\alpha = 1.4$ $b = 0.3$	2000		2	$D = 1.286$ $\sigma = 1.273$ $\nu = 1.229$	$\nu = 1.231$	$D = 1.272$
			3	$D = 1.244$ $\sigma = 1.261$ $\nu = 1.221$	$\nu = 1.231$	$\sigma = 1.258$ $\nu = 1.224$
Lorentz equation $R = 28$ $\sigma = 10$ $b = 8/3$	4000	0.2	3	$D = 1.778$ $\sigma = 2.028$ $\nu = 2.049$	$\nu = 2.069$	
			4	$D = 1.821$ $\sigma = 2.079$ $\nu = 2.106$	$\nu = 2.120$	$\nu = 2.05$
			5	$D = 1.802$ $\sigma = 2.072$ $\nu = 2.071$	$\nu = 2.156$	

applied to a small data set, we performed numerical tests on three classical examples, namely, the logistic map

$$x_{n+1} = \lambda x_n(1 - x_n), \tag{7}$$

the Hénon map

$$\begin{aligned} x_{n+1} &= y_n + 1 - ax_n^2, \\ y_{n+1} &= bx_n, \end{aligned} \tag{8}$$

and the Lorenz equation

$$\begin{aligned} \dot{x} &= \sigma(y - x), \\ \dot{y} &= -y - xz + Rx, \\ \dot{z} &= xy - bz, \end{aligned} \tag{9}$$

with discrete points at times with spacing $\delta t = 0.2$.

In order to simulate the real experimental situation, we deliberately choose a small data set and consider only one variable. The phase recovering technique is used. To our experience, the time lag does not influence the calculation results very much. We take a lag of one entry throughout the paper. The corresponding results are shown in Table I and Figs. 1–3, which show clearly that the box-counting algorithm gives the correlation dimension ν more accurately than those with the Grassberger–Procaccia algorithm in all three examples, though it only needs about 20% computation amount of that of the latter one as it is indicated in Table II. From these results we argue that box-counting as

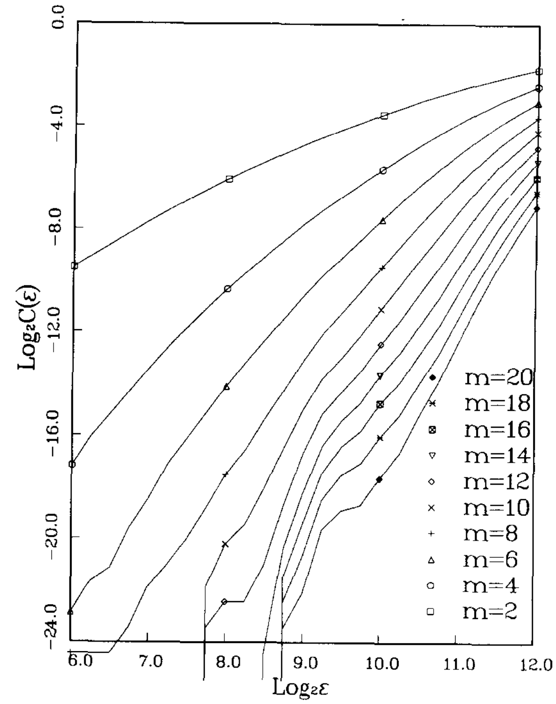

FIG. 5. Correlation dimension of pulsar 0950 + 08 calculated with Grassberger–Procaccia algorithm by using 4000 entries.

applied to the correlation dimension of an experimental data set is in general less time-consuming and at least as accurate as the Grassberger–Procaccia algorithm, although the latter one is superior to applying box-counting to calculate the capacity dimension.

The systematic errors of the Grassberger–Procaccia algorithm are recently considered numerically in [12]. It is possible to do a similar analysis for the box-counting algorithm.

The box-counting algorithm has been applied to calculate the correlation dimension of EEG signals, and a reasonably good result was obtained in short computer times [13]. In the following we present the results of treating pulsar intensity data.

The intensity data of pulsar 0950 + 08 were obtained in April 1976 at the Arecibo Observatory, which are divided into three sets with 2000, 4800, and 4000 pulses, respectively. These data were analyzed and discussed in [14],

TABLE II

CPU Time (seconds) for Box-Counting and the Grassberger–Procaccia Algorithm Using a VAX-7800

Number of iterations	500	1000	2000	4000
Box-counting	6	13	43	132
Grassberger–Procaccia	25	73	191	822

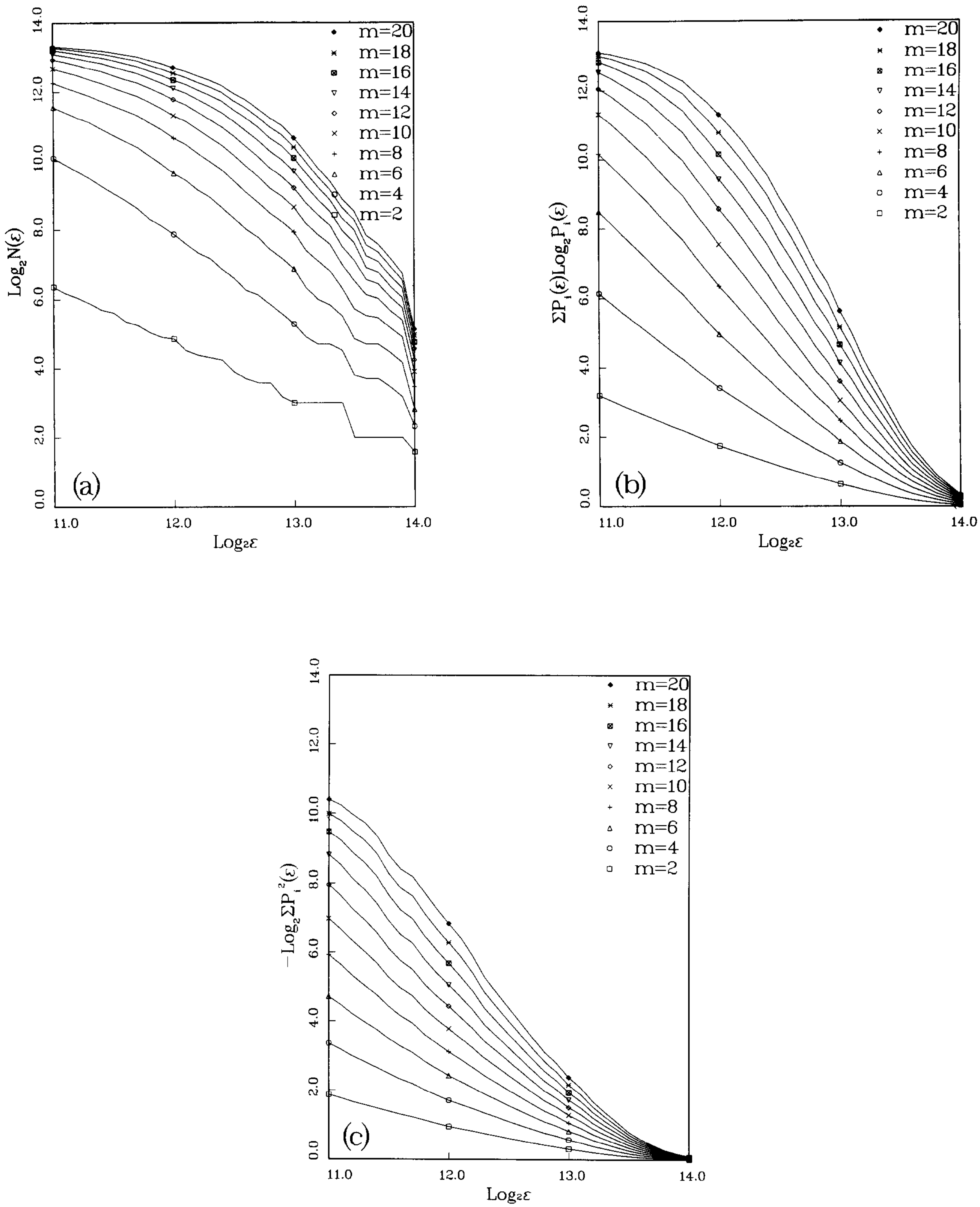

FIG. 6. Generalized dimensions of pulsar 0950 + 08 calculated with box-counting by using 10,800 entries: (a) capacity, (b) information dimension; (c) correlation dimension.

however, not from the viewpoint of a dimensional analysis. The three data sets exhibit similar behaviour, but the second set seems to have less noise.

With the box-counting algorithm we calculated generalized dimensions of the pulsar. By doing so, the pulse intensity data were first transformed to integers to simplify the programming. The maximum pulse intensity is less than 2^{15}, i.e., a box with side length 2^{15} will include all points. From Fig. 4 we estimate capacity $D \approx 4.0$, information dimension $\sigma \approx 6.0$ and correlation dimension $v \approx 4.5$ with an embedding dimension ≈ 14. As we mentioned before, the capacity calculated with box-counting will be smaller than it should be, but we can trust the information dimension and especially the correlation dimension more or less. The correlation integral obtained by using the Grassberger–Procaccia algorithm are shown in Fig. 5; the nearly straight part of the curves are not very well parallel, but we can estimate a correlation dimension ≈ 4.5 and an embedding dimension ≈ 14 from this picture.

We also calculated generalized dimensions of the combination of all data sets (10,800 entries), and the results are shown in Fig. 6. One does not find a remarkable difference between corresponding pictures in Fig. 4 and Fig. 6. This means a data set with some 4000 entries is enough for the box-counting algorithm in this case.

For a complex system such as a pulsar, the best thing one can expect might be to find an approximate low-dimensional strange attractor and to estimate its fractal dimension and the minimum embedding dimension with a reasonable accuracy. Since two algorithms give similar results of the dimension values, we expect they have a certain meaning, and this deserves further investigation.

In conclusion we point out, that it is strongly recommended to calculate the correlation dimension of an experimental data set with the box-counting algorithm, since it requires very little computer time and gives enough good results for the practical purpose.

ACKNOWLEDGMENTS

This work was supported by DE-FG02-87ER13740 and AFOSR-87-0122. We thank Jim Cordes for providing the data and the National Astronomy and Ionosphere Center, which operates the Arecibo Observatory under contract with the NSF.

REFERENCES

1. H. Froehling, J. P. Crutchfield, J. D. Farmer, N. H. Packard, and R. S. Shaw, *Physica D* **3**, 605 (1981).
2. N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw, *Phys. Rev. Lett.* **45**, 712 (1980).
3. F. Takens, in *Lecture Notes in Math.*, Vol. 898 (Springer-Verlag, Berlin, 1981, p. 336.
4. A. M. Fraser, *Physica D* **34**, 391 (1989).
5. P. Grassberger, *Phys. Lett. A* **97**, 224 (1983).
6. P. Grassberger and I. Procaccia, *Phys. Rev. Lett.* **50**, 346 (1983).
7. P. Grassberger and I. Procaccia, *Physica D* **9**, 189 (1983).
8. P. Grassberger and I. Procaccia, *Physica D* **13**, 34 (1984).
9. F. H. Ling, *Phys. Lett. A* **138**, 25 (1989).
10. J. Theiler, *Phys. Rev. A* **36**, 4456 (1987).
11. H. S. Greenside *et al.*, *Phys. Rev. A* **25**, 3453 (1982).
12. M. Möller, W. Lange, F. Mitschke, N. B. Abraham, and U. Hübner, *Phys. Lett. A* **138**, 176 (1989).
13. F. H. Ling, N. Xu, and J. H. Xu, *Neural Networks* **1** (Suppl. 1), 108 (1988).
14. T. H. Hankins and J. M. Cordes, *Astrophys. J.* **249**, 241 (1981).