Journal of Statistical Physics, Vol. 62, Nos. 3/4, 1991

Some Results on the Behavior and Estimation of the Fractal Dimensions of Distributions on Attractors

C. D. Cutler[1]

Received February 16, 1990; final July 18, 1990

The strong interest in recent years in analyzing chaotic dynamical systems according to their asymptotic behavior has led to various definitions of fractal dimension and corresponding methods of statistical estimation. In this paper we first provide a rigorous mathematical framework for the study of dimension, focusing on pointwise dimension $\sigma(x)$ and the generalized Renyi dimensions $D(q)$, and give a rigorous proof of inequalities first derived by Grassberger and Procaccia and Hentschel and Procaccia. We then specialize to the problem of statistical estimation of the correlation dimension ν and information dimension σ. It has been recognized for some time that the error estimates accompanying the usual procedures (which generally involve least squares methods and nearest neighbor calculations) grossly underestimate the true statistical error involved. In least squares analyses of ν and σ we identify sources of error not previously discussed in the literature and address the problem of obtaining accurate error estimates. We then develop an estimation procedure for σ which corrects for an important bias term (the local measure density) and provides confidence intervals for σ. The general applicability of this method is illustrated with various numerical examples.

KEY WORDS: Information dimension; correlation dimension; fractal dimension; fractal measures; dynamical systems; attractors.

1. INTRODUCTION

Over the past decade there has been much interest in the asymptotic behavior of dynamical systems, particularly in the case of systems exhibiting "chaotic" behavior. A feature of many chaotic systems is the apparent existence of a "strange" or "fractal" attracting set (or *attractor*) on to which

[1] Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

0022-4715/91/0200-0651$06.50/0 © 1991 Plenum Publishing Corporation

the trajectories of the system eventually settle. Considerable effort has gone into attempts to describe and quantify attractors and this has led to the development of several definitions of *fractal dimension*. The early papers of Farmer *et al.*,[28] Grassberger and Procaccia,[30,31] and Hentschel and Procaccia[37] introduced many of these notions and additionally provided conjectures (not all correct) and some proofs concerning the relationships among them. The recent interest in "multifractal" theory, in large part initiated by the paper of Halsey *et al.*,[35] has led to an increased understanding of the role and meaning of the various dimension concepts in dynamical systems, providing, in particular, a link between the variation of pointwise dimension in small neighborhoods and the generalized Renyi dimensions. Nonetheless, the standard approach to definitions and proofs (which generally consists of covering the attractor by a grid of nonoverlapping cubes) requires certain regularity assumptions (often not explicitly stated), such as existence of a limit (as cube size goes to zero) which does not depend on the choice of grid. One goal of this paper is to present and clarify the different definitions of dimension in a systematic and mathematically rigorous way, developing the relevant properties and inequalities under minimal assumptions. This is the content of Section 2. The remainder of the paper specializes to a study of the correlation dimension ν and the information dimension σ (to be defined in Section 2), focusing in particular on the statistical techniques used to estimate these quantities. Thus, from the point of view of statistics, our attention is concentrated on measure-dependent concepts of dimension; that is, concepts which require the existence of a natural probability distribution $\mathscr{m}$ on the attractor. [Typically, for a subset B of the attractor, $\mathscr{m}(B)$ represents the long-run proportion of time a system trajectory spends in B.] Measure-dependent notions of dimension incorporate information about the actual dynamics of the system and are often related to relevant dynamical quantities such as Lyapunov exponents and singularity spectra (e.g., refs. 35, 45, 46, and 63). We will not attempt any solution to the problem of estimating the fractal dimension (taken here to be the Hausdorff dimension or perhaps the capacity) of the attractor itself viewed solely as a geometric object. In general, estimation of a geometric fractal dimension is a particularly hard problem. The computational difficulties involved with box-counting algorithms are well known,[33] but it is important to realize that the difficulty in estimating geometric quantities from data goes deeper than simply the problem of finding an efficient algorithm. Treating data merely as a geometric set of points ignores the inherent probabilistic nature of its structure. A finite data set will reveal only certain regions of an attractor (those with highest probability) and geometric dimensions may be rendered inestimable (as remarked by a referee) except in special circumstances.

Some difficulties and solutions concerning estimation of geometric dimensions are presented in Tricot *et al.*,[62] Dubuc *et al.*,[22] and Taylor and Taylor.[56] For a history of the different definitions of dimension and corresponding statistical techniques we refer the reader to Farmer *et al.*,[28] Badii and Politi,[1] Eckmann and Ruelle,[23] Mayer-Kress,[40] and Paladin and Vulpiani.[46]

In Section 3 we critique the typical least squares procedure performed when estimating dimension. (Our analysis is done in terms of the correlation dimension, but the basic conclusions are equally applicable to least squares procedures for estimating σ.) We show that distributions on attractors typically give rise to a spatial correlation structure that features a *wandering intercept*; this can lead to failure of the slopes to converge to the desired parameter v as the radius $r \rightarrow 0$. Furthermore, unlike the simple linear regression model, the observed proportions at different radii constitute correlated observations with unequal variances, and hence the usual errors associated with a simple linear regression (such as the mean squared deviation from the least squares line) are totally inadequate to describe the actual statistical error in the problem. Denker and Keller[21] made an important contribution by developing the theory leading to the asymptotic joint distribution and covariance structure of the observed proportions for time series data taken from dynamical systems with adequate mixing properties. It is a theory which appears to be underused in practice, so in Section 3.2 we provide the explicit form of the asymptotic covariance matrix as well as explicit estimators for the components of the matrix. The technique is seen to perform extremely well in a numerical study of the Kaplan–Yorke map.

At the beginning of Section 4 we note that generally the problem of a wandering intercept is more severe for least squares procedures involving σ; this is due in part to the variability from point to point caused by the "fractalness" of the measure, and in part to the effects of the local measure density at different points x. The remainder of Section 4 deals with the development of a nearest neighbor technique for obtaining confidence intervals for σ which corrects for the bias due to the local measure density. This technique arose out of rigorous results[19,20] on the asymptotic behavior of nearest neighbors from different classes of distributions. Naturally no technique can be expected to work well in all cases, but we present several examples to illustrate the general adaptability of this method to a variety of dynamical systems. Not surprisingly, the data requirements increase rapidly as the true underlying dimension increases. We expect that modifications of this method (perhaps by incorporating a set of nearest neighbors) will lead to more efficient use of the data and improved estimates.

2. DIMENSION DEFINITIONS, PROPERTIES, AND INEQUALITIES

We begin first with a general discussion of dimension and probability distributions before specializing to the case of dynamical systems. The reader is likely familiar with the definition of *Hausdorff dimension* (which has long been associated with the concept of fractals) and perhaps less familiar with the much more recent notion of *packing dimension.*[57,58,61] While Hausdorff dimension measures the size of a set by considering optimal coverings by sets of small diameter, packing dimension measures size by considering optimal *packings* of the set by small disjoint balls centered at points of the set. (For reasonable sets the Hausdorff and packing dimensions will agree.) The notion of packing dimension has turned out to be the missing key in many questions concerning dimension, and its introduction allows us to state various theorems in a very complete way. We therefore give precise definitions of both Hausdorff and packing dimensions below.

Let $\mathscr{M}$ be a metric space with metric ρ, and let $E \subseteq \mathscr{M}$. By a *δ-covering* of E we will mean a countable collection $\{S_k\}_k$ of subsets of $\mathscr{M}$ with diameter $\rho(S_k) \leqslant \delta$ such that $E \subseteq \bigcup_k S_k$. We will define a *$\delta$-packing* of E to be a countable disjoint collection $\{B(x_k, r_k)\}_k$ of closed balls centered at points $x_k \in E$ with radius $r_k \leqslant \delta/2$. (Note that a δ-packing need not be a covering of E and in fact generally will not cover, due to the disjointness constraint.)

The (α, δ)-outer Hausdorff measure $H^\alpha_\delta(E)$ is defined to be

$$H^\alpha_\delta(E) = \inf \left\{ \sum_k \rho(S_k)^\alpha \;\middle|\; \{S_k\}_k \text{ is a } \delta\text{-covering of } E \right\} \tag{2.1}$$

while the (α, δ)-premeasure $P^\alpha_\delta(E)$ is defined by

$$P^\alpha_\delta(E) = \sup \left\{ \sum_k \rho(B_k)^\alpha \;\middle|\; \{B_k\}_k \text{ is a } \delta\text{-packing of } E \right\} \tag{2.2}$$

The Hausdorff measure $H^\alpha(E)$ is then obtained by letting the covering size δ go to zero:

$$H^\alpha(E) = \lim_{\delta \to 0} H^\alpha_\delta(E) \tag{2.3}$$

while the packing measure $P^\alpha(E)$ is constructed by a two-stage procedure as the packing size δ tends to zero:

$$\bar{P}^\alpha(E) = \lim_{\delta \to 0} P^\alpha_\delta(E) \tag{2.4}$$

and then

$$P^{\alpha}(E) = \inf\left\{\sum_k \bar{P}^{\alpha}(E_k) \;\middle|\; E = \bigcup_k E_k\right\} \tag{2.5}$$

The second step is necessary to ensure that P^{α} is a (countably-. additive) measure, a property already satisfied by H^{α}. (If we used $\bar{P}^{\alpha}$ to define a dimensional index, the result would be a packing analogue of *capacity*. We will not discuss capacity, but a definition can be found in Farmer *et al.*[28])

The Hausdorff dimension $\dim(E)$ and the packing dimension $\operatorname{Dim}(E)$ are then defined by

$$\begin{aligned} \dim(E) &= \inf\{\alpha \mid H^{\alpha}(E) = 0\} = \sup\{\alpha \mid H^{\alpha}(E) = \infty\} \\ \operatorname{Dim}(E) &= \inf\{\alpha \mid P^{\alpha}(E) = 0\} = \sup\{\alpha \mid P^{\alpha}(E) = \infty\} \end{aligned} \tag{2.6}$$

The properties of Hausdorff measure and dim are discussed in Rogers[50] and Falconer,[27] while the properties of packing measure and Dim are developed in Taylor and Tricot[57,58] and Saint Raymond and Tricot.[52] It is a known result that in a separable metric space $P^{\alpha}(E) \geqslant H^{\alpha}(E)$ and, as a consequence, $\operatorname{Dim}(E) \geqslant \dim(E)$. Taylor[55] has suggested that the term "fractal" be reserved for sets E which satisfy the condition $\dim(E) = \operatorname{Dim}(E)$, thereby forcing some degree of regularity in the structure of the set. Standard self-similar sets will meet this requirement. [The reader should not confuse this condition with the much more stringent restriction requiring equality of $\dim(E)$ and capacity $C(E)$. The set of rational numbers Q satisfies $\dim(Q) = \operatorname{Dim}(Q) = 0$ while $C(Q) = 1$.] A set E for which $\dim(E) \neq \operatorname{Dim}(E)$ might be considered hopelessly irregular. We will see shortly that, from a statistical viewpoint, equality of the Hausdorff and packing dimensions is exactly what is needed to obtain good theoretical results.

We now wish to connect the above concepts of dimension (of sets) with probability distributions living on the space. If m is a probability measure on the Borel sets of $\mathcal{M}$, the distribution of m-mass with respect to the Hausdorff and packing dimensions can be described by related probability measures m_{H} and m_{P} defined on the Borel sets of $[0, \infty]$ via

$$\begin{aligned} m_{\mathrm{H}}([0, \alpha]) &= \sup\{m(D) \mid \dim(D) \leqslant \alpha\} \\ m_{\mathrm{P}}([0, \alpha]) &= \sup\{m(D) \mid \operatorname{Dim}(D) \leqslant \alpha\} \end{aligned} \tag{2.7}$$

m_{H} (with the alternate notation $\hat{m}$) is discussed in Cutler,[14] while m_{P} is introduced in Cutler.[18] We always have $m_{\mathrm{H}}([0, \alpha]) \geqslant m_{\mathrm{P}}([0, \alpha])$ because of the inequality $\dim(E) \leqslant \mathrm{Dim}(E)$. We will say that $\mathscr{m}$ is *dimension regular* if $\mathscr{m}_{\mathrm{H}} = \mathscr{m}_{\mathrm{P}}$. Most measures arising in practice appear to satisfy this regularity condition (it takes work to build examples where $\mathscr{m}_{\mathrm{H}} \neq \mathscr{m}_{\mathrm{P}}$). In general, all manner of distributions are possible candidates for $\mathscr{m}_{\mathrm{H}}$ and $\mathscr{m}_{\mathrm{P}}$; to each probability distribution ω defined on $[0, N]$ corresponds infinitely many probability measures $\mathscr{m}$ on $\mathbb{R}^N$ satisfying $\mathscr{m}_{\mathrm{H}} = \mathscr{m}_{\mathrm{P}} = \omega$. Under certain circumstances (such as in the case of smooth ergodic dynamical systems, to be discussed below) the situation simplifies enormously; both $\mathscr{m}_{\mathrm{H}}$ and $\mathscr{m}_{\mathrm{P}}$ collapse to point masses. We say that $\mathscr{m}$ is of *exact* Hausdorff (respectively, packing) dimension if there exists $\alpha \geqslant 0$ such that $\mathscr{m}_{\mathrm{H}} = \delta_\alpha$ (respectively, $\mathscr{m}_{\mathrm{P}} = \delta_\alpha$), where δ_α denotes the unit mass at α. Note that $\mathscr{m}_{\mathrm{H}} = \delta_\alpha$ if and only if $\mathscr{m}$ can be supported on some set of Hausdorff dimension α but has no mass on any set of smaller dimension. (It is possible, even in smooth ergodic systems, for $\mathscr{m}$ to be of exact Hausdorff dimension α but of exact packing dimension β, where $\alpha \neq \beta$. Hence dimension regularity is not automatic. We will say more about this later.) If $\mathscr{m}_{\mathrm{H}} = \mathscr{m}_{\mathrm{P}} = \delta_\sigma$, we call the common value σ the *information dimension* of $\mathscr{m}$. (Otherwise, we say the information dimension does not exist.)

Remark. Two other approaches to defining information dimension exist, both of which coincide with ours if $\mathscr{m}$ is suitably regular and exact-dimensional (in both the Hausdorff and packing sense). It is not uncommon to see the definition $\sigma^* = \inf\{\dim(E) \mid \mathscr{m}(E) = 1\}$, which considers only the Hausdorff dimension of the largest set needed to support $\mathscr{m}$. The Rényi dimension approach, perhaps most common, is a little different (see Farmer *et al.*[28] or Hentschel and Procaccia[37]) and actually corresponds to computing an average. Specifically, covering the attractor by the minimum possible number $N(\varepsilon)$ of cubes C_k of side length ε, the information dimension is often taken to be

$$\sigma^{**} = \lim_{\varepsilon \to 0} \frac{I(\varepsilon)}{\log 1/\varepsilon} \qquad \text{where} \quad I(\varepsilon) = -\sum_{k=1}^{N(\varepsilon)} \mathscr{m}(C_k) \log \mathscr{m}(C_k) \tag{2.8}$$

assuming that this limit exists and is well defined. See the last comment in the proof of Theorem 2.2 for details on the relationship between σ^{**} and our approach. A connection between σ^{**} and entropy is made via the Shannon–McMillan–Breiman theorem (see Billingsley[8]). The advantages to our approach are, first, that it does not require any (hidden) assumptions on the measure structure, and, second, that the concept of "information dimension" is allowed to exist precisely in those situations where it has a clear-cut meaningful interpretation and is estimable from data (this connects with certain statistical properties of data, discussed in Section 4.)

The dimension structure of $\mathscr{m}$ can be explored in even greater detail by considering the pointwise mappings of $\mathscr{M}$ into $[0, \infty]$ defined by

$$\begin{aligned} \sigma_{\mathrm{H}}(x) &= \liminf_{r \to 0} \frac{\log \mathscr{m}(B(x, r))}{\log r} \\ \sigma_{\mathrm{P}}(x) &= \limsup_{r \to 0} \frac{\log \mathscr{m}(B(x, r))}{\log r} \end{aligned} \tag{2.9}$$

Defining the preimage sets D_{H}^{α} and D_{P}^{α} by

$$D_{\mathrm{H}}^{\alpha} = \{x \mid \sigma_{\mathrm{H}}(x) \leqslant \alpha\}, \qquad D_{\mathrm{P}}^{\alpha} = \{x \mid \sigma_{\mathrm{P}}(x) \leqslant \alpha\} \tag{2.10}$$

it can be shown[15,18] that

$$\dim(D_{\mathrm{H}}^{\alpha}) \leqslant \alpha, \qquad \mathrm{Dim}(D_{\mathrm{P}}^{\alpha}) \leqslant \alpha \tag{2.11}$$

$$\mathscr{m}_{\mathrm{H}}([0, \alpha]) = \mathscr{m}(D_{\mathrm{H}}^{\alpha}), \qquad \mathscr{m}_{\mathrm{P}}([0, \alpha]) = \mathscr{m}(D_{\mathrm{P}}^{\alpha}) \tag{2.12}$$

Equation (2.12) implies that if a point X is chosen randomly according to the measure $\mathscr{m}$, $\sigma_{\mathrm{H}}(X)$ and $\sigma_{\mathrm{P}}(X)$ may be regarded as random variables on $\mathscr{M}$ with distributions $\mathscr{m}_{\mathrm{H}}$ and $\mathscr{m}_{\mathrm{P}}$, respectively. [Hence $\mathscr{m}_{\mathrm{H}}$ and $\mathscr{m}_{\mathrm{P}}$ are completely specified by knowledge of $\sigma_{\mathrm{H}}(X)$ and $\sigma_{\mathrm{P}}(X)$.] It is clear that $\mathscr{m}$ is dimension regular if and only if $\sigma_{\mathrm{H}}(x) = \sigma_{\mathrm{P}}(x)$ $\mathscr{m}$-a.s. [The "only if" part follows from the fact that we always have the inequality $\sigma_{\mathrm{P}}(x) \geqslant \sigma_{\mathrm{H}}(x)$.] We also see that $\mathscr{m}$ is of exact Hausdorff (respectively, packing) dimension α if and only if $\sigma_{\mathrm{H}}(x) = \alpha$ $\mathscr{m}$-a.s. [respectively, $\sigma_{\mathrm{P}}(x) = \alpha$ $\mathscr{m}$-a.s.].

Remark. Relationships between the pointwise limit

$$\lim_{r \to 0^+} \frac{\log \mathscr{m}(B(x, r))}{\log r}$$

and the dimension behavior of $\mathscr{m}$ have been known or conjectured for some time. Young[63] proved that if $\mathscr{m}(S) > 0$ and for every $x \in S$

$$\underline{\delta} \leqslant \liminf_{r \to 0^+} \frac{\log \mathscr{m}(B(x, r))}{\log r} \leqslant \limsup_{r \to 0^+} \frac{\log \mathscr{m}(B(x, r))}{\log r} \leqslant \bar{\delta}$$

then $\underline{\delta} \leqslant \dim(S) \leqslant \bar{\delta}$. The first results of this kind were apparently proved by Billingsley.[6,7] Tricot[60] expanded on some of Billingsley's results and also discussed the concept of regularity. In Cutler[14,15] it is shown explicitly that the "lim inf" in (2.9) describes the distribution of $\mathscr{m}$-mass with respect to Hausdorff dimension; the relationship between "lim sup" and packing dimension is developed in Cutler.[18]

Note that Eq. (2.11) provides some information about the structure of the preimages D_{H}^{α} and D_{P}^{α}, specifically giving upper bounds on the Hausdorff/packing dimensions of these sets. This is connected with the current theory of multifractals, where the chief point of interest is the behavior of the function $f(\alpha) = \dim(D_0^{\alpha})$, where $D_0^{\alpha} = \{x \mid \sigma_{\mathrm{H}}(x) = \sigma_{\mathrm{P}}(x) = \alpha\}$. Most examples studied so far in multifractal theory possess sufficient self-similarity or regularity in the measure structure that there is no real distinction between Hausdorff and packing dimensions [and $f(\alpha)$ exhibits smooth properties as a function of α]. See refs. 3, 4, 11, 13, 32, 35, 46, and 49.

We define the *average Hausdorff dimension* $\bar{\sigma}_{\mathrm{H}}$ and *average packing dimension* $\bar{\sigma}_{\mathrm{P}}$ by

$$\begin{aligned} \bar{\sigma}_{\mathrm{H}} &= E(\sigma_{\mathrm{H}}(X)) = \int \sigma_{\mathrm{H}}(x)\, m(dx) = \int \alpha\, m_{\mathrm{H}}(d\alpha) \\ \bar{\sigma}_{\mathrm{P}} &= E(\sigma_{\mathrm{P}}(X)) = \int \sigma_{\mathrm{P}}(x)\, m(dx) = \int \alpha\, m_{\mathrm{P}}(d\alpha) \end{aligned} \tag{2.13}$$

If $\bar{\sigma}_{\mathrm{H}} = \bar{\sigma}_{\mathrm{P}}$ (which will occur if and only if m is dimension regular), it might be reasonable to call the common value $\bar{\sigma}$ the *average information dimension* of m. Obviously, if m is additionally exact-dimensional, we have $\bar{\sigma} = \sigma$.

The mappings $\sigma_{\mathrm{H}}(x)$ and $\sigma_{\mathrm{P}}(x)$ enable us to study the pointwise scaling behavior of m in small neighborhoods. The generalized Rényi dimensions, introduced in Hentschel and Procaccia,[37] enable us to study global scaling properties of m. We will use an L^q-norm approach to defining these quantities (Chapter 3 of Rudin[51] provides the basic theory of L^q spaces.) A parallel development of the generalized dimensions, using grids of cubes, can be found in Beck.[3]

For each $r > 0$ let $V_r(x) = m(B(x, r))$. Then, for $q \neq 0$, we define the L^q norm of V_r by

$$\|V_r\|_q = E(V_r(X)^q)^{1/q} = \left[\int V_r(x)^q\, m(dx)\right]^{1/q} \tag{2.14}$$

Such norms are usually only defined for $q > 0$, but we will allow $q < 0$ with the understanding that $\|V_r\|_q = 0$ if $E(V_r(X)^q) = \infty$ for $q < 0$. Since $0 \leqslant V_r \leqslant 1$, we clearly have $\|V_r\|_q \leqslant 1$ when $q > 0$. However, $E(V_r(X)^q)$ may explode for $q < 0$. A simple example is the probability measure m with density function $g(x) = e^{-x}$ for $x > 0$. It is easy to determine that $E(V_r(X)^q) = \infty$ whenever $q \leqslant -1$.

The following theorem summarizes some well-known results on L^q norms. We include the proof of the basic inequalities for completeness.

Theorem 2.1. The norms $\{\|V_r\|_q\}_q$ $(q \neq 0)$ increase as a function of q. Furthermore, $\lim_{q \to \infty} \|V_r\|_q = \|V_r\|_\infty$ (where $\|V_r\|_\infty$ denotes the essential supremum of V_r), and $\lim_{q \to 0^+} \|V_r\|_q = \exp(E(\log V_r(X)))$.

Proof. The norm inequalities are a simple consequence of Jensen's inequality, which states that if ϕ is a convex function, then $E(\phi(X)) \geqslant \phi(E(X))$. Noting that $\phi(t) = t^\alpha$ is convex for $\alpha < 0$ or $\alpha \geqslant 1$ (and concave for $0 < \alpha < 1$), we obtain

$$\begin{aligned} E(V_r(X)^q) = E((V_r(X)^p)^{q/p}) &\geqslant E(V_r(X)^p)^{q/p} \qquad \text{if} \quad 0 < p < q \text{ or } p < 0 < q \\ \text{but} \quad &\leqslant E(V_r(X)^p)^{q/p} \qquad \text{if} \quad p < q < 0 \end{aligned}$$

We then see that $\|V_r\|_p \leqslant \|V_r\|_q$ whenever $p < q$. Refer to Rudin[51] for information on the other statements in the theorem. ▮

We now define the lower and upper qth *moment dimensions* $(q \neq 0)$ by

$$\begin{aligned} m^-(q) &= \liminf_{r \to 0} \frac{\log \|V_r\|_q}{\log r} \\ m^+(q) &= \limsup_{r \to 0} \frac{\log \|V_r\|_q}{\log r} \end{aligned} \tag{2.15}$$

If $m^-(q) = m^+(q)$, we denote the common value by $m(q)$. Note that $m(q) = \infty$ is possible. The *generalized Rényi dimensions* $D(q)$ [under the assumption $m(q)$ exists] are then defined by

$$D(q) = m(q-1) \qquad \text{for} \quad q \neq 1 \tag{2.16}$$

Remark. When defining the Rényi dimensions using grids of cubes (as done in Beck[3]) it is more natural to structure the q-scale so that $q = 1$ (rather than $q = 0$) is the undefined point, which produces Eq. (2.16).

We now present some inequalities involving the moment dimensions and the average Hausdorff/packing dimensions which generalize some of the inequalities in Hentschel and Procaccia[37] and Beck.[3]

Theorem 2.2. Let $\mathscr{m}$ be a probability measure defined on the Borel sets of a compact set $\mathscr{M} \subseteq \mathbb{R}^N$. Then:

(a) $\{m^-(q)\}$ and $\{m^+(q)\}$, $q \neq 0$, decrease as functions of q.

(b) $m^-(p) \geqslant \bar{\sigma}_H$ and $\bar{\sigma}_P \geqslant m^+(q)$ whenever $p < 0 < q$.

(c) If $\mathscr{m}$ is dimension regular, then $m^-(p) \geqslant \bar{\sigma}_H = \bar{\sigma}_P = \bar{\sigma} \geqslant m^+(q)$ whenever $p < 0 < q$.

Proof. Part (a) is immediate from (2.15) and Theorem 2.1. To see the first inequality in (b), note that

$$\begin{aligned}
\bar{\sigma}_{\mathrm{H}} &= E(\sigma_{\mathrm{H}}(X)) \\
&= E\left(\liminf_{r\to 0} \frac{\log V_r(X)}{\log r}\right) \\
&\leqslant \liminf_{r\to 0} E\left(\frac{\log V_r(X)}{\log r}\right) && \text{by Fatou's lemma} \\
&= \liminf_{r\to 0} E\left(\frac{\log V_r(X)^p}{\log r^p}\right) \\
&\leqslant \liminf_{r\to 0} \frac{\log E(V_r(X)^p)}{\log r^p} && \text{by Jensen's inequality when } p<0 \\
&= m\ \ (p)
\end{aligned}$$

To obtain the second inequality in (b), write

$$\begin{aligned}
\bar{\sigma}_{\mathrm{P}} &= E(\sigma_{\mathrm{P}}(X)) \\
&= E\left(\limsup_{r\to 0} \frac{\log V_r(X)}{\log r}\right) \\
&\geqslant \limsup_{r\to 0} E\left(\frac{\log V_r(X)}{\log r}\right) && \text{if the functions are uniformly integrable} \\
&= \limsup_{r\to 0} E\left(\frac{\log V_r(X)^q}{\log r^q}\right) \\
&\geqslant \limsup_{r\to 0} \frac{\log E(V_r(X)^q)}{\log r^q} && \text{by Jensen's inequality when } q>0 \\
&= m^+(q)
\end{aligned}$$

In Appendix A we prove that the functions involved are always uniformly integrable, so the above argument is rigorous. Part (c) now follows from (b) and the fact that dimension regularity is equivalent to the equality $\bar{\sigma}_{\mathrm{H}} = \bar{\sigma}_{\mathrm{P}}$. Further note that under this equality we can exchange the limit and expectation operator to obtain

$$\bar{\sigma} = E\left(\lim_{r\to 0} \frac{\log V_r(X)}{\log r}\right) = \lim_{r\to 0} E\left(\frac{\log V_r(X)}{\log r}\right) = \sigma^{**}$$

using a more careful definition of σ^{**} than that given in (2.8). ▮

It is obviously of interest to determine the conditions under which $m^{-}(q)=m^{+}(q)$ and under which we have continuity across the boundary point $q=0$. If $\mathscr{m}$ is dimension regular and continuity exists across $q=0$, it is then reasonable to define $m(0)=D(1)=\bar{\sigma}$.

In this paper we will chiefly be interested in the statistical estimation of σ (under the assumption that $\mathscr{m}$ is of exact Hausdorff/packing dimension σ) and of the *correlation dimension* $\nu=m(1)=D(2)$. Although a member of the Rényi dimensions, correlation dimension was introduced earlier, as a separate concept, by Grassberger and Procaccia.[30] It is clear that the choice $q=1$ plays a special role among the L^q norms. Letting $C(r)$ denote the probability that two random, independent points (chosen according to $\mathscr{m}$) are no more than distance r apart, we have

$$\begin{aligned}
C(r) &= \iint I_{A(r)}(x, y)\, \mathscr{m}(dx)\, \mathscr{m}(dy) \\
&\qquad [\text{where } A(r)=\{(x, y) \mid \|x-y\| \leqslant r\}] \\
&= E(\mathscr{m}(B(X, r))) \qquad \text{the mean or expected mass in a random ball of radius } r \\
&= \|V_r\|_1 \qquad \text{the } L^1 \text{ norm of } V_r
\end{aligned} \tag{2.17}$$

$C(r)$ is often called the *spatial correlation integral.* Note then that

$$\nu=\lim_{r \to 0} \frac{\log \|V_r\|_1}{\log r}=\lim_{r \to 0} \frac{\log C(r)}{\log r}$$

provided this limit exists. Thus, ν describes the asymptotic (as $r \to 0$) scaling behavior of the average mass in a ball of radius r. This should be contrasted with σ, which is concerned with the scaling behavior of the mass at individual points in the space. Because ν is the more easily measured quantity (the prime reason for its introduction by Grassberger and Procaccia[30]) it has become a popular choice for "dimension" among experimentalists. However, it is important to note that there is no intrinsic relationship between ν and the size (in terms of Hausdorff/packing dimension) of the smallest sets capable of supporting $\mathscr{m}$. The inequality $\sigma \geqslant \nu$ is immediate from (c) of Theorem 2.2, but, as has been discussed by Ott *et al.*,[43] Beck,[3] and Cutler,[17] it is possible to construct very plausible examples where ν and σ are very far apart. There are smooth ergodic dynamical systems in $\mathbb{R}^N$ where the invariant measure $\mathscr{m}$ satisfies $\sigma=N$ while ν is arbitrarily close to zero. (See Example 3.3 in this paper for a system of this form.) It has come to be recognized that correlation dimension, and the Rényi dimensions as a whole, provide information not about the size of supporting sets, but about nonuniformity in the measure m across its support. If m has a bounded Radon–Nikodym derivative (i.e., density) with respect to the uniform measure across its support [so $m(B) = \int_B g(x)\,dx$, where $|g(x)| \leqslant K$], then we will observe $\sigma = m(q)$ for all $q > 0$. However, if m possesses singularities (as is the typical case), then the Rényi dimensions will differ from σ and among themselves (to an extent which reflects the degree and nature of the singularities.) This connects with the multifractal theory mentioned earlier.

We now turn our attention specifically to the invariant distributions associated with dynamical systems. Throughout the paper we consider the case where $\mathcal{M}$ is a compact subset of $\mathbb{R}^N$, m is a probability measure on the Borel sets of $\mathcal{M}$, and $T: \mathcal{M} \to \mathcal{M}$ is a Borel-measurable mapping which preserves m (i.e., $m = mT^{-1}$). We further assume that the system $(\mathcal{M}, T, m)$ is ergodic. Hence, under repeated iterations of T the time averages of bounded functions along the orbits of T converge, for m-almost all x, to the corresponding space averages. That is,

$$\lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} f(T^j(x)) = \int f(z)\, m(dz) \quad m\text{-a.s.}$$

for all bounded measurable functions $f: \mathcal{M} \to \mathbb{R}$.

Remark. In practice, initial conditions are very often selected according to the natural Lebesgue measure on a compact manifold $\mathcal{M}$ and it is then observed or conjectured that the empirical measures

$$\frac{1}{n} \sum_{j=1}^{n} \delta_{T^j(x)}$$

converge weakly to an ergodic measure m on an attractor. This scenario demands much more than simple ergodicity of m, since typically m is singular with respect to Lebesgue measure. Such measures m which attract orbits corresponding to almost all initial conditions (in the Lebesgue sense) are often called *Bowen–Ruelle* measures, and their existence has been proven rigorously in the case of Axiom A attractors.[10] We will assume, regardless of the manner in which initial conditions were selected, that the system has evolved through a sufficient number of iterations to be regarded as invariant and ergodic. Experimentally this may mean ignoring several hundred or several thousand initial iterates in order to allow transients time to die out. In situations where a physical phenomenon cannot be reproduced at will (such as in the case of climatological data), one cannot afford the luxury of ignoring observations; however, it is not unreasonable in many such cases to assume that the system has already been evolving for some time and has achieved a probabilistic equilibrium.

Cutler[15] proved that if $(\mathcal{M}, T, m)$ is an ergodic system where T is sufficiently smooth, then m is of exact Hausdorff dimension α for some $\alpha \geqslant 0$. The analogous result holds also for packing dimension.[18] Sufficient smoothness is guaranteed if T is differentiable at m-almost all x, a condition easily seen to be met by the maps considered in this paper. We do point out, however, that Ledrappier and Misiurewicz[39] have shown that dimension regularity is not an automatic consequence of smoothness and ergodicity, so that we may observe $m_{\mathrm{H}} = \delta_\alpha$ and $m_{\mathrm{P}} = \delta_\beta$, where $\alpha \neq \beta$. While Ledrappier and Misiurewicz[39] as well as Young[62] have established certain sufficient conditions for dimension regularity in one- and two-dimensional maps, there does not yet appear to be a known general minimum criterion for dimension regularity. However, for most of the maps considered in this paper, equality of the pointwise Hausdorff and packing dimensions can be verified directly. In Section 4 associated statistical properties are discussed and applied to the problem of estimating the common value σ.

Another property which we will often require for purposes of estimation is that of mixing. Recall that a system $(\mathcal{M}, T, m)$ is said to be *mixing* if, for all Borel sets $A \subseteq \mathcal{M}$ and $B \subseteq \mathcal{M}$,

$$\lim_{n \to \infty} m(A \cap T^{-n}(B)) = m(A)\, m(B) \tag{2.18}$$

The mixing property allows us to treat observations sampled sufficiently far apart on a trajectory as possessing a certain degree of stochastic independence, so that a statistical analysis may be carried out. It may sometimes be necessary to assume a stronger form of mixing, such as the existence of a sequence of *mixing coefficients* $\alpha(n) \downarrow 0$ satisyfing

$$\sup_{A,B} |m(A \cap T^{-n}(B)) - m(A)\, m(B)| \leqslant \alpha(n) \tag{2.19}$$

The mixing coefficients may also be required to approach 0 at a specified rate. We do not go into the details of the mixing properties of maps here, since Denker and Keller[21] have given a good discussion of this and the role of mixing in estimation. Most of the maps examined in this paper can be shown to exhibit some form of the mixing property.

3. LEAST SQUARES ANALYSES OF THE CORRELATION DIMENSION

In the following we assume that $W_1, \ldots, W_n$ are n observations taken along some orbit of an ergodic dynamical system possessing good mixing properties. The observations need not be successive; in fact, whether we choose to sample successively or at spaced intervals should depend on the degree of mixing believed to be present. Denker and Keller[21] show that sufficient mixing is present in many dynamical systems to justify the sampling of successive observations (at least within the context of the present problem). We note that n is actually the *effective* sample size in this procedure. The true total sample size includes all discarded observations (which may be considerable in number).

We assume that the correlation dimension

$$v = \lim_{r \to 0^+} \frac{\log C(r)}{\log r}$$

exists (with $0 < v < \infty$), suggesting an asymptotic linear relationship between $\log r$ and $\log C(r)$. Let $\mathbf{r} = (r_1, ..., r_m)$ be a fixed vector of radii $r_1 > r_2 > \cdots > r_m > 0$. Let $x_i = \log r_i$ and define $z_i = \log C(r_i)$. Note that $x_1, ..., x_m$ and $z_1, ..., z_m$ are fixed (nonrandom) quantities. $z_1, ..., z_m$ are unknown and may be regarded as parameters whose values depend on the underlying distribution $\mathscr{m}$. $C(r)$ is estimated from the data $W_1, ..., W_n$ by the sample proportion $C_n(r)$ of pairs of observations that are no more than the distance r apart. That is,

$$C_n(r) = \binom{n}{2}^{-1} \sum_{j<k}\sum I_{\{\|W_j - W_k\| \leqslant r\}} \tag{3.1}$$

Set $y_i = \log C_n(r_i)$ and note that y_i is a point estimate of z_i. The standard method for estimating v has been to plot y_i vs. x_i and take the slope b of the least squares line through the data pairs $(x_1, y_1), ..., (x_m, y_m)$ as the point estimate of v (assuming that the line shows a good fit to the data). Note that the only role of the sample size n of the original observations $W_1, ..., W_n$ is in the quality of the estimate y_i of z_i.

We will discuss two distinct aspects of the problem of estimating v from the pairs $(x_1, y_1), ..., (x_m, y_m)$. The first concerns our ability to actually identify v in a linear model (this is the problem which we refer to as *the wandering intercept*). The second deals with the fact that the random variables $y_1, ..., y_m$ are statistically correlated and possess distinct variances c_{ii} (in contrast to the simple linear regression model, which assumes uncorrelated observations with common variance c; we refer the reader to Myers.[41]) Before proceeding to these discussions, we state a convergence result, based on the theory of U-statistics, proved by Denker and Keller.[21] This result is the main justification for the estimation techniques employed in this section, and may be considered a substitute for the usual assumptions that accompany a simple linear regression model. For the technical details concerning the smoothness and mixing assumptions in Theorem 3.1 we refer the reader to Denker and Keller's paper.

Theorem 3.1 (Denker and Keller). Let $(\mathcal{M}, T, m)$ be a smooth ergodic dynamical system with good mixing properties. Let $W_1, ..., W_n$ be a sequence of observations from an orbit of this system. Then:

(a) For each fixed $r > 0$, $C_n(r)$ and $\log C_n(r)$ converge with probability 1 (wp1) to $C(r)$ and $\log C(r)$, respectively, as $n \to \infty$.

(b) For each vector $C(\mathbf{r}) = (C(r_1), ..., C(r_m))$ there exists a nonnegative-definite $m \times m$ matrix $\mathsf{U} = (u_{ij})$ such that the sequence of normalized vectors $n^{1/2}(C_n(\mathbf{r}) - C(\mathbf{r})) \to N_m(\mathbf{0}, \mathsf{U})$ in distribution as $n \to \infty$, where $N_m(\mathbf{0}, \mathsf{U})$ denotes an m-dimensional Gaussian distribution with mean vector $\mathbf{0}$ and covariance matrix U. The analogous result holds for $n^{1/2}[\log C_n(\mathbf{r}) - \log C(\mathbf{r})]$ with covariance matrix V, where $v_{ij} = u_{ij}/C(r_i)\, C(r_j)$. First and second moments converge correspondingly. ∎

3.1. The Wandering Intercept

Consider first the ideal case where the actual parameter values $C(r_1), ..., C(r_m)$ are known (no need for data or estimation). Note that we can always write the identity

$$\begin{aligned}\log C(r) &= \log\left(\frac{C(r)}{r^\nu}\right) + \nu \log r \\ &= \alpha(r) + \nu \log r \qquad (3.2)\end{aligned}$$

where the intercept $\alpha(r)$ depends on r. Corresponding to the vector $\mathbf{r} = (r_1, ..., r_m)$ we thus have the following system of m equations in the $m + 1$ unknowns $\alpha(r_1), ..., \alpha(r_m)$, and ν:

$$\begin{matrix} z_1 = \alpha(r_1) + \nu x_1 \\ \vdots \quad \vdots \quad \vdots \\ z_m = \alpha(r_m) + \nu x_m \end{matrix} \qquad (3.3)$$

This reveals the basic problem of *nonidentifiability* of ν in this model. The assumption that the limit

$$\nu = \lim_{r \to 0^+} \frac{\log C(r)}{\log r}$$

exists does not guarantee convergence of the intercept $\alpha(r)$ to some constant α as $r \to 0$. In fact, we will argue that in the generic case $\alpha(r)$

wanders or *oscillates* as $r \to 0$, sometimes even diverging to $\pm\infty$. It will be seen that in this situation the choice of vector $\mathbf{r}$ (and the way in which components of $\mathbf{r}$ are made to approach 0) plays a significant role. [This contrasts with an assumption often made for purposes of estimation; namely that, for small r, $C(r)$ follows an exact power law, e.g., Takens.[54]]

Now consider the case where the parameter values $z_1,..., z_m$ are estimated from the data values $y_1,..., y_m$. In view of (3.3) and Theorem 3.1, the corresponding statistical model is

$$\begin{array}{l} y_1 = \alpha(r_1) + vx_1 + e_1 \\ \vdots \quad\;\; \vdots \qquad\;\; \vdots \quad\;\; \vdots \\ y_m = \alpha(r_m) + vx_m + e_m \end{array} \tag{3.4}$$

where $e_1,..., e_m$ are error random variables. Assuming n is sufficiently large, Theorem 3.1 says that the vector $n^{1/2}(e_1,..., e_m)$ has approximately an $N_m(\mathbf{0}, \mathrm{V})$ distribution for some covariance matrix V. We will need to consider the nature of this covariance matrix later. At this stage the important point is that $y_1,..., y_m$ are asymptotically (as $n \to \infty$) unbiased estimates of $z_1,..., z_m$.

Now consider m points $(u_1, w_1),..., (u_m, w_m)$ in $\mathbb{R}^2$. By minimizing the sum of squares of deviations $\sum_{i=1}^m (w_i - a - bu_i)^2$ over all possible choices of a and b, one obtains the least squares line $w = a + bu$, where $b = S_{uw}/S_{uu}$, $a = \bar{w} - b\bar{u}$, $S_{uu} = \sum_{i=1}^m (u_i - \bar{u})^2$, and

$$S_{uw} = \sum_{i=1}^{m} (u_i - \bar{u})(w_i - \bar{w}) = \sum_{i=1}^{m} (u_i - \bar{u})\, w_i$$

Here $\bar{u}$ and $\bar{w}$ denote the arithmetic means of, respectively, the u values and w values.

We will let $v(\mathbf{r})$ denote the slope S_{xz}/S_{xx} of the least squares line through the parameter pairs $(x_1, z_1),..., (x_m, z_m)$. Note that $v(\mathbf{r})$ does not necessarily coincide with v except in the special case $\alpha(r_1) = \cdots = \alpha(r_m)$. The points $(x_1, z_1),..., (x_m, z_m)$ need not lie on a straight line, and even if they do, the slope $v(\mathbf{r})$ of that line may differ from v. [In this last case we say that $C(\mathbf{r})$ exhibits *spurious scaling behavior* at $\mathbf{r}$.] We will let $b(\mathbf{r})$ denote the slope S_{xy}/S_{xx} of the least squares line through the data pairs $(x_1, y_1),..., (x_m, y_m)$. Note that $b(\mathbf{r})$ is a function of the original sample $W_1,..., W_n$ (as are $y_1,..., y_m$), but for simplicity we suppress "n" in the notation. We obtain the following result.

Theorem 3.2. Assume that the hypotheses of Theorem 3.1 are satisfied and the normalized error variables $n^{1/2}(e_1,..., e_m)$ asymptotically follow an $N_m(\mathbf{0}, \mathrm{V})$ distribution, $\mathrm{V} = (v_{ij})$. Then $\lim_{n\to\infty} b(\mathbf{r}) = v(\mathbf{r})$ with probability 1, $\lim_{n\to\infty} E(b(\mathbf{r})) = v(\mathbf{r})$, and $n^{1/2}(b(\mathbf{r}) - v(\mathbf{r}))$ asymptotically follows a Gaussian distribution with mean 0 and variance

$$S_{xx}^{-2} \sum_{i=1}^{m} \sum_{j=1}^{m} (x_i - \bar{x})(x_j - \bar{x})\, v_{ij}$$

Proof. Write

$$b(\mathbf{r}) = S_{xx}^{-1} \sum_{i=1}^{m} (x_i - \bar{x})\, y_i$$

Since $\lim_{n\to\infty} y_i = z_i$ wp1, it follows that

$$\lim_{n\to\infty} b(\mathbf{r}) = S_{xx}^{1} \sum_{i=1}^{m} (x_i - \bar{x})\, z_i = v(\mathbf{r}) \quad \text{wp1}$$

Convergence of the first moments occurs similarly. Now since the asymptotic distribution of $n^{1/2}(e_1,\ldots, e_m)$ is $N_m(\mathbf{0}, \mathrm{V})$ and $b(\mathbf{r})$ is a linear function of $y_1,\ldots, y_m$ (and hence of $e_1,\ldots, e_m$), the rest of the theorem follows. ▌

The above theorem shows that $b(\mathbf{r})$ is a consistent estimator of $v(\mathbf{r})$. This is only useful in the present problem if $v(\mathbf{r})$ is close to v. One might hope that, in practice, a value of $v(\mathbf{r})$ which is very different from v will result in a least squares line which fits the data poorly, thereby removing any temptation to treat $v(\mathbf{r})$ as the correlation dimension. While certainly a poor fit will occur in some such cases, it is not uncommon to observe false scaling behavior over certain $\mathbf{r}$ vectors. This may pose a genuine difficulty when dealing with a real finite data set. It is less likely to be a problem in a laboratory situation where unlimited data can be generated and many (and longer) $\mathbf{r}$ vectors plotted and compared [see (c) of Theorem 3.3].

Substituting $\alpha(r_i) + vx_i$ for z_i in the expression for $v(\mathbf{r})$, we see that

$$\begin{aligned} v(\mathbf{r}) &= S_{xx}^{1} \sum_{i=1}^{m} (x_i - \bar{x})\, \alpha(r_i) + vS_{xx}^{1} \sum_{i=1}^{m} (x_i - \bar{x})\, x_i \\ &= v + S_{xx}^{1} \sum_{i=1}^{m} (x_i - \bar{x})\, \alpha(r_i) \end{aligned}$$

Consequently, the *parameter error* or *systematic error* corresponding to $\mathbf{r}$ is

$$d(\mathbf{r}) = v(\mathbf{r}) - v = S_{xx}^{-1} \sum_{i=1}^{m} (x_i - \bar{x})\, \alpha(r_i) \tag{3.5}$$

It is clear that the value $d(\mathbf{r})$ generally depends on the choice of $\mathbf{r}$ and the behavior of the intercept function at the points $r_1, \ldots, r_m$. The following theorem examines the way in which r should be chosen in order to force $d(\mathbf{r}) \to 0$.

Theorem 3.3. Assume

$$v = \lim_{r \to 0^+} \frac{\log C(r)}{\log r}$$

exists. Then:

(a) $\lim_{r \to 0^+} [\alpha(r)/\log r] = 0$. This is the most that can be said in general concerning the convergence behavior of the intercept function (and is in fact equivalent to the hypothesis of the theorem).

(b) Suppose there exists a finite constant α such that $\lim_{r \to 0^+} \alpha(r) = \alpha$. [In this case we say that $C(r)$ exhibits *true scaling behavior*.] Let $\mathbf{r} = (r_1, \ldots, r_m)$ be fixed and let $0 < s < 1$. If we shrink the vector $\mathbf{r}$ by the geometric factor s, then the parameter error goes to 0. That is, $\lim_{k \to \infty} d(s^k \mathbf{r}) = 0$. This is generally (although not always) false if $\alpha(r)$ fails to converge.

(c) Let $r_0 > 0$ and $0 < s < 1$. For each m define the mth vector $\mathbf{r}_m = (s^m r_0, s^{m+1} r_0, \ldots, s^{2m-1} r_0)$. Then $\lim_{m \to \infty} d(\mathbf{r}_m) = 0$.

Proof. Part (a) is immediate from the definition of $\alpha(r)$. To see (b), set $x_i = \log r_i$ and $x_{k,i} = \log s^k r_i$. Then $x_{k,i} - \bar{x}_k = x_i - \bar{x}$ for each k and i, where $\bar{x}_k$ and $\bar{x}$ are, respectively, the arithmetic means of $x_{k,1}, \ldots, x_{k,m}$ and $x_1, \ldots, x_m$. Consequently, $S_{x_k x_k} = S_{xx}$ for each k, giving

$$d(s^k \mathbf{r}) = S_{xx}^{-1} \sum_{i=1}^{m} (x_i - \bar{x})\, \alpha(s^k r_i)$$

From this it is clear that if $\alpha(r) \to \alpha$, then $d(s^k \mathbf{r}) \to 0$ as $k \to \infty$. To prove (c), note that, for $x_{m,j} = \log s^{m-1+j} r_0$ and $\bar{x}_m = (1/m) \sum_{j=1}^{m} x_{m,j}$, we have

$$\begin{aligned} S_{x_m x_m} &= \sum_{j=1}^{m} (x_{m,j} - \bar{x}_m)^2 = (\log s)^2 \sum_{j=1}^{m} \left(j - \frac{m+1}{2} \right)^2 \\ &= (\log s)^2 \left(\frac{m(m+1)(m-1)}{12} \right) \end{aligned}$$

Since $\alpha(r)/\log r \to 0$, there exists $\varepsilon(m) \to 0$ such that $|\alpha(r)| \leqslant \varepsilon(m)\, |\log r|$ for all r in $(0, s^m r_0]$. Hence from the Cauchy–Schwarz inequality and (3.5) we obtain

$$\begin{aligned}
|d(\mathbf{r}_m)| &\leqslant S^{-1}_{x_m x_m} S^{1/2}_{x_m x_m} \left(\sum_{j=1}^{m} \alpha(s^{m-1+j} r_0)^2 \right)^{1/2} \\
&\leqslant S^{-1/2}_{x_m x_m} \left(\sum_{j=1}^{m} (\varepsilon(m) \log s^{m-1+j} r_0)^2 \right)^{1/2} \\
&\leqslant \varepsilon(m)\, S^{-1/2}_{x_m x_m} \left(\sum_{j=1}^{m} (\log s^{2m-1} r_0)^2 \right)^{1/2} \\
&\leqslant \varepsilon(m)\, K[m(m+1)(m-1)]^{-1/2}\, m^{3/2}
\end{aligned}$$

for some constant K which does not depend on m. Hence we see that $d(\mathbf{r}_m) \to 0$ as $m \to \infty$. ▌

In practice a limit to the applicability of the above asymptotic theory is of course the original effective sample size n. For small r the quality of the estimate $C_n(r)$ of $C(r)$ is poor. (This will be discussed later in greater detail.) However, even when estimation is restricted to a certain r range, Theorem 3.3 can prove useful in detecting the presence of a wandering intercept. In the following example we show that the uniform distribution across the Cantor set in $[0, 1]$ gives rise to a wandering intercept. We then illustrate the way in which this behavior can be detected numerically.

Example 3.1. The Cantor Distribution. Let K denote the standard Cantor set in $[0, 1]$ constructed by removing middle thirds. Note that we may write $K = \bigcap_{m=1}^{\infty} K_m$, where $K_m = \bigcup_{i=1}^{2^m} K_{m,i}$ is the finite disjoint union of 2^m intervals of length 3^{-m}. The Cantor distribution $\mathscr{m}_K$ is the unique probability measure on $[0, 1]$ which assigns $\mathscr{m}_K(K_{m,i}) = 2^{-m}$ for each m and i. [A dynamical system corresponding to K and $\mathscr{m}_K$ is given by the ternary shift map $T(x) = 3x$ (mod 1), where initial conditions are selected randomly and uniformly from K. It is well known that the system $(K, T, \mathscr{m}_K)$ is ergodic with strong mixing properties.] Since $\mathscr{m}_K$ is uniform across K, we obtain $\nu = \sigma = \log 2/\log 3$. (The value $\log 2/\log 3$ can easily be deduced from the approach of Hentschel and Procaccia,[37] utilizing the self-similarity properties of K.)

To show that the intercept $\alpha(r)$ of $\mathscr{m}_K$ wanders, we will exhibit two sequences $u_m \to 0$ and $w_m \to 0$ such that $\alpha(u_m) = 0$ for each m while

$$\alpha(w_m) = \log\left(\frac{5/2}{5^{\log 2/\log 3}}\right) \cong -0.0992$$

for each m. We take $u_m = 3^{-m}$ and $w_m = 5 \cdot 3^{-m}$. Now, since the intervals $K_{m,1}, \ldots, K_{m,2^m}$ are separated by empty intervals of length at least 3^{-m}, it is easy to see that

$$C(u_m) = \mathscr{m}_K \times \mathscr{m}_K(\{(x, y) \mid |x - y| \leqslant 3^{-m}\}) = 2^m 2^{-m} 2^{-m} = 2^{-m}$$

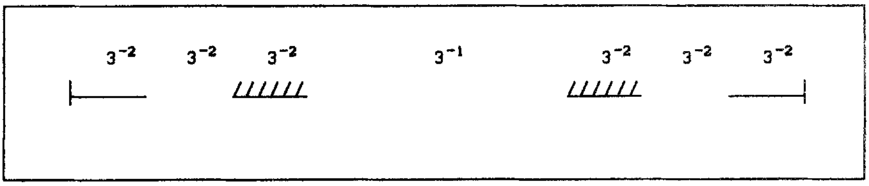


Fig. 1. The Cantor set at level $m = 2$. The shaded interval on the left is I_1. The shaded interval on the right is I_2.

Consequently,

$$\alpha(u_m) = \log(C(u_m)/u_m^{\nu}) = \log(2^{\ m}/2^{\ m}) = \log 1 = 0$$

A little more work is required in the case of w_m. Consider first $m = 2$ and refer to Fig. 1. The set $\{(x, y) \mid |x - y| \leqslant 5 \cdot 3^{\ 2}\}$ is the disjoint union of the set $\{(x, y) \mid |x - y| \leqslant 3^{\ 1}\}$ with $I_1 \times I_2$ and $I_2 \times I_1$, where I_1 and I_2 are the two shaded intervals indicated in Fig. 1. Note in particular that $3^{\ 2} + 3^{\ 1} + 3^{\ 2} = 5 \cdot 3^{\ 2}$. Hence we obtain

$$\begin{aligned} C(5 \cdot 3^{\ 2}) &= C(3^{\ 1}) + C(\text{pairs in } I_1 \times I_2 \text{ and } I_2 \times I_1) \\ &= 2^{\ 1} + 2 \cdot (2^{\ 2} \cdot 2^{\ 2}) = 5 \cdot 2^{\ 3} \end{aligned}$$

Now consider $m = 3$ and refer to Fig. 2. The set $\{(x, y) \mid |x - y| \leqslant 5 \cdot 3^{\ 3}\}$ is the disjoint union of the sets $\{(x, y) \mid |x - y| \leqslant 3^{\ 2}\}$, $I_1 \times I_2$, $I_2 \times I_1$, $I_3 \times I_4$, and $I_4 \times I_3$, where I_1 and I_2 are the shaded intervals of length 3^{-3} on the left side of Fig. 2 and I_3 and I_4 are the shaded intervals of length 3^{-3} on the right side of Fig. 2. We obtain

$$\begin{aligned} C(5 \cdot 3^{-3}) &= C(3^{-2}) + C(\text{pairs in } I_1 \times I_2, I_2 \times I_1, I_3 \times I_4, I_4 \times I_3) \\ &= 2^{-2} + 2^2 \cdot (2^{-3} \cdot 2^{-3}) = 5 \cdot 2^{-4} \end{aligned}$$

Carrying on in this manner, we obtain the general result $C(w_m) = 5 \cdot 2^{-(m+1)}$. This shows that the intercept function fails to converge, since if we calculate along the sequence $\{w_m\}$, we obtain

$$\alpha(w_m) = \log\left(\frac{C(w_m)}{w_m^{\nu}}\right) = \log\left(\frac{5/2}{5^{\log 2/\log 3}}\right)$$

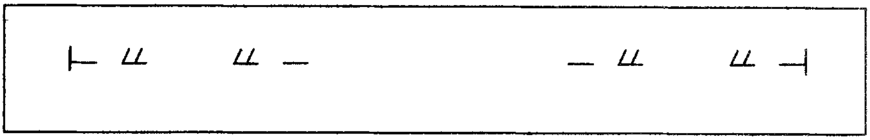


Fig. 2. The Cantor set at level $m = 3$. The shaded intervals on the left are I_1 and I_2. The shaded intervals on the right are I_3 and I_4.

The effect of the wandering intercept can be seen by considering the vectors $\mathbf{r}_m = (u_{m-1}, u_m)$ and $\mathbf{s}_m = (w_m, u_m)$, $m = 2, 3,\ldots$. Since $\alpha(u_{m-1}) = \alpha(u_m)$, it follows from (3.5) that $d(\mathbf{r}_m) = 0$ for each m. Hence $\mathbf{r}_m$ provides the correct value $v(\mathbf{r}_m) = v \cong 0.63093$ for each m. However, since

$$\alpha(w_m) = \log\left(\frac{5/2}{5^{\log 2/\log 3}}\right) \neq \alpha(u_m) = 0$$

$\mathbf{s}_m$ produces a constant parameter error

$$d(\mathbf{s}_m) = (\log 5)^{-1} \log\left(\frac{5/2}{5^{\log 2/\log 3}}\right) \approx -0.06161.$$

Hence we have $v(\mathbf{s}_m) = v + d(\mathbf{s}_m) \cong 0.56932$ for each m. This illustrates the remark made in (b) of Theorem 3.3 that the systematic error generally does not vanish when shrinking a vector by a geometric factor in the wandering intercept case. Numerically this can be observed by comparing the slopes $b(\mathbf{r}_m)$ and $b(\mathbf{s}_m)$ of the straight lines through each of the pairs $(\log \mathbf{r}_m, \log C_n(\mathbf{r}_m))$ and $(\log \mathbf{s}_m, \log C_n(\mathbf{s}_m))$ for various values of m. In Table I we present the results of a numerical study for $m = 4, 5, 6$. We carried out five simulations, each producing a data set of $n = 2000$ (sequential) observations from $\mathscr{m}_K$. For each m, each data set was analyzed twice, first using $\mathbf{r}_m$ and then using $\mathbf{s}_m$.

Comparing the two columns in Table I, we see clear evidence of the

Table I. The Effect of the Wandering Intercept: Slopes over Two Vectors $\mathbf{r}_m$ and $\mathbf{s}_m$ for the Uniform Distribution on the Cantor Set

	$b(\mathbf{r}_m)$	$b(\mathbf{s}_m)$
$m = 3$	0.63121	0.56812
	0.63105	0.56704
	0.63199	0.56840
	0.63273	0.57479
	0.63133	0.57701
$m = 4$	0.62452	0.55912
	0.63046	0.56887
	0.63148	0.57805
	0.62708	0.55751
	0.62587	0.55439
$m = 5$	0.62436	0.55750
	0.63019	0.56944
	0.63246	0.57845
	0.62486	0.55707
	0.62680	0.55418

wandering intercept. Note that in order to determine conclusively that certain differences between vectors are due to actual systematic errors and not simply statistical error (due to sampling and estimation), it is usually necessary to have more than one data set so that repetition is possible. (For example, in Table I, five different data sets were used. It is clear that the variation within each column—the statistical error—is much less than the variation between columns, indicating a real parameter difference.) However, even in the absence of repetition (a situation very likely to be encountered in practice) the evaluation of the slope at different levels of two or more vectors can be informative. In Table I we see that following any one data set reveals a consistency in the slopes over the different levels of $\mathbf{r}_m$ and $\mathbf{s}_m$ that suggests a systematic error. In practice we can look for such behavior and move to mitigate it by employing (c) of Therem 3.3. This may require obtaining additional data so that estimates at smaller values of r can be obtained. Ideally over these longer vectors we would then observe convergence of the slopes toward a common value v.

This indicates that the presence of a wandering intercept represents the generic case where dynamical systems are concerned. The justification for making such a claim is the current popular belief that most attractors are Cantorian at least in some direction. The analysis carried out in the preceding example indicates that inexact scaling behavior is probably a fundamental characteristic of distributions over Cantor-type sets. Note that the methods described above enable us to look for this feature, and subsequently provide some information about the geometric structure of the system in addition to determining v. It is important to realize that while the exhibited parameter error is small in Example 3.1, we should expect to be able to observe exaggerated errors by modifying the geometry of the underlying Cantor set (especially in higher dimensions) and considering nonuniform measures across the set. It is interesting to speculate whether the apparently poor scaling behavior of the Zaslavskii map[31] is due to failure of the correlation dimension to exist or to excessive oscillations in the intercept function. The work of Termonia and Alexandrowicz[59] suggests that the associated measure may not be exact-dimensional.

We conclude this section with two examples. These examples illustrate the very different behavior that can arise even when the attracting set is not a fractal. They will be revisited in Section 4 when we consider the information dimension.

Example 3.2. The Logistic Map. The logistic map $T(x) = 4x(1-x)$ on $[0, 1]$ is a standard example of an ergodic dynamical system possessing an invariant distribution $\mathscr{m}$ which is equivalent to Lebesgue measure. Hence the attractor of the system is $[0, 1]$. The probability density function $g(x)$ of m is given by $g(x) = \pi^{-1}[x(1-x)]^{-1/2}$. It can be shown that $\nu = \sigma = 1$ here, but this example represents a borderline case for the correlation dimension in the sense that a density function with a sharper singularity at 0 would result in $\nu < 1$. (See Example 3.3.) This borderline aspect is reflected in the behavior of the intercept function, which actually diverges to ∞ as $r \to 0$. It can be seen that $C(r) = O(r \log 1/r)$ and hence $\alpha(r) = O(\log \log 1/r)$. The result is that the method of letting $\mathbf{r} \to 0$ described in (c) of Theorem 3.3 produces a sequence $\nu(\mathbf{r}_m)$ which converges very slowly to $\nu = 1$. The inexact scaling here has been discussed by Grassberger and Procaccia,[31] who suggest embedding the observations in higher-dimensional space; see Packard *et al.*[44] and Takens.[53] (This method is often used when attempting to estimate ν for a higher-dimensional attractor based only on a single-variable time series; grouping of the data into vectors of length d and embedding into $\mathbb{R}^d$ is carried out before least squares is performed. The manner of grouping the data and the statistical correlations between successive embeddings are of course additional sources of error variation which should be accounted for in a proper statistical analysis; however, we consider this topic to be outside the scope of the present paper.) Grassberger and Procaccia state that in some cases this embedding procedure appears to reduce the systematic error. The difficulty in applying this or any other technique which works well only in certain situations lies in our general inability to identify (based only on data) those instances in which a particular method will perform best. We note that a recent paper of Ramsey and Yuan[48] examines some of the questions involved with embeddings and estimation of ν in general.

We present the results of a simulation study using a variant of Theorem 3.3(c) to illustrate the gradual convergence of $\nu(\mathbf{r})$ toward 1 for the logistic map. We generated five data sets with 10,000 observations apiece, and analyzed each over the following four vector levels:

$$\mathbf{r}_1 = 0.09\,(s^1 s^2 s^3 s^4), \qquad \mathbf{r}_2 = 0.09\,(s^2 s^3 s^4 s^5 s^6)$$

$$\mathbf{r}_3 = 0.09\,(s^3 s^4 s^5 s^6 s^7 s^8), \qquad \mathbf{r}_4 = 0.09\,(s^4 s^5 s^6 s^7 s^8 s^9 s^{10})$$

where $s = 1/3$. Table II shows the mean intercept and mean slope (over the

Table II. Slope and Intercept of the Least Squares Line for the Logistic Map over Four Different Vector Levels

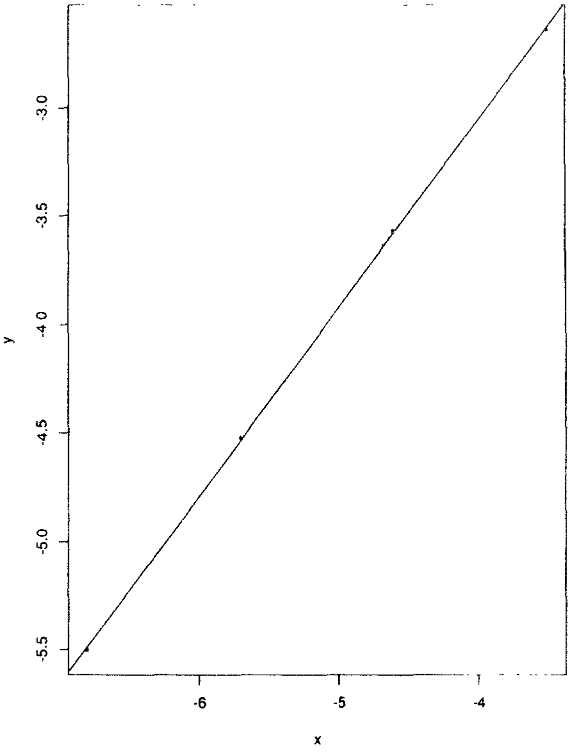
	Level 1	Level 2	Level 3	Level 4
Intercept	0.402	0.531	0.663	0.822
Slope	0.863	0.885	0.903	0.921

five data sets) at each vector. Figure 3 shows the scatter plot and least squares line at each vector for one of the data sets. Note that excellent scaling behavior is observed at each level (which illustrates the danger in accepting a good fit based on a single vector). By considering the four vectors we observe the shifting slope and shifting intercept as r decreases. This demonstrates the usefulness of an iterative procedure to check the validity of conclusions.

Example 3.3. Cutler[17] examined the family of mappings $\{T_\gamma\}$, $0<\gamma<1/2$, where

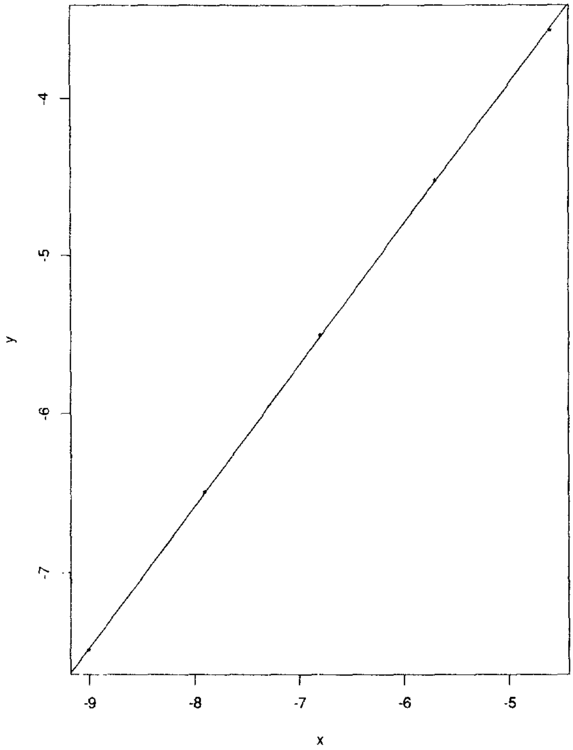
$$T_\gamma(x)=\begin{cases} 2x & 0\leqslant x<1/2 \\ (2x-1)^{1/\gamma} & 1/2\leqslant x\leqslant 1 \end{cases}$$

Corresponding to T_γ is a unique ergodic measure m_γ which is equivalent to Lebesgue measure on $[0, 1]$ (and hence $\sigma=1$). The density function $g(x)$



(a)

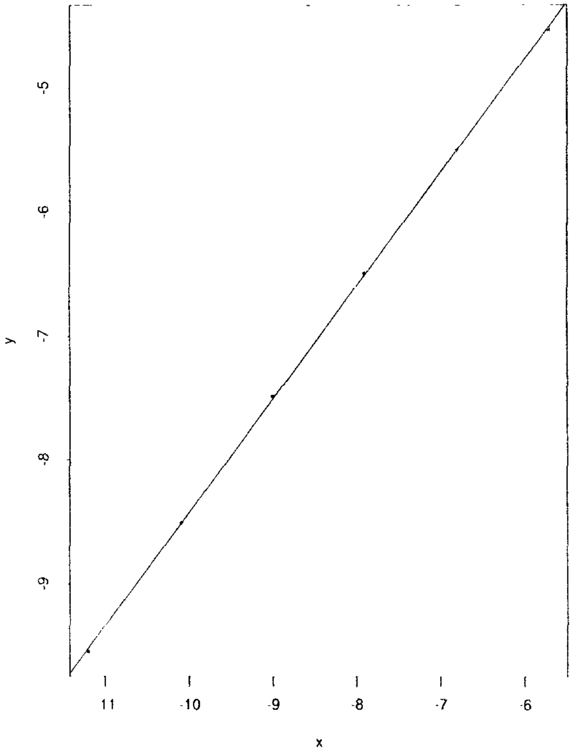
Fig. 3. Scatter plots and least squares lines for the logistic map over four different vector levels.



(b)

Fig. 3. (*Continued*)

of m_γ satisfies the inequalities $Cx^{\gamma-1} \leqslant g(x) \leqslant Kx^{\gamma-1}$ for some finite positive constants C and K (which depend on γ). It is shown that $\nu = 2\gamma$ and that the intercept $\alpha(r)$ is bounded between two finite constants. For $\gamma = 1/3$ this gives $\nu \approx 0.667$, and we found this value to be more accessible numerically than its counterpart for the logistic map. As in Example 3.2, five data sets were analyzed at the vector $\mathbf{r}_4$ and produced a mean slope of 0.677. (It should be noted, however, that the individual slopes were quite variable, ranging between 0.621 and 0.701.) This seems to indicate that the presence of sharp singularities in the density does not in itself imply poor convergence; the crucial point is whether or not the singularity represents a boundary case where scaling behavior is very inexact. However, note that for small γ we expect very slow mixing in this particular system [as iterates spend a great deal of time just doubling in the interval $(0, \varepsilon)$] and to obtain quality estimates it may become necessary to sample far apart on an orbit.



(c)

Fig. 3. (*Continued*)

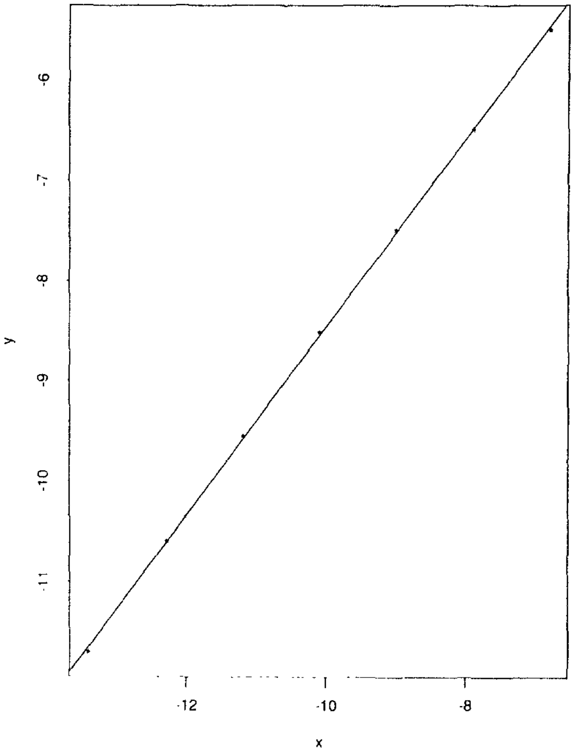
3.2. Estimating the Variability of $b(\mathbf{r})$

The previous section dealt with the systematic error $v(\mathbf{r}) - v$. Here we consider the statistical behavior of the estimate $b(\mathbf{r})$. Our main tools are Theorems 3.1 and 3.2 of the previous section, which describe the asymptotic behavior of $b(\mathbf{r})$. Letting c_{ij} denote the covariance between y_i and y_j, for sufficiently large n (from Theorems 3.2 and 3.1 we expect $c_{ij} \approx v_{ij}/n$) the variance θ of $b(\mathbf{r})$ is given by:

$$\theta = S_{xx}^{-2} \sum_{i=1}^{m} \sum_{j=1}^{m} (x_i - \bar{x})(x_j - \bar{x})\, c_{ij}$$

$$\approx S_{xx}^{-2} \sum_{i=1}^{m} \sum_{j=1}^{m} (x_i - \bar{x})(x_j - \bar{x})\, v_{ij}/n$$

Below we derive the asymptotic covariance matrix $\mathsf{U} = (u_{ij})$ of the normalized proportions $n^{1/2} C_n(\mathbf{r})$ in the special case that the observations $W_1, \ldots, W_n$ are independent. The matrix V then follows by Theorem 3.1.



(d)

Fig. 3. (*Continued*)

Theorem 1 of Denker and Keller[21] can be used to obtain U (and we will appeal to their result for the general case of correlated observations), but it is instructive to derive the matrix in a simple case. We see that the asymptotic covariance structure is determined by the variability of the distribution $\mathscr{m}$ from point to point across its support. Let W denote a random observation from $\mathscr{m}$ and define

$$\begin{aligned}\tau(r_1, r_2) &= \operatorname{Cov}(\mathscr{m}(B(W, r_1)), \mathscr{m}(B(W, r_2)))\\ &= \int \mathscr{m}(B(w, r_1))\, \mathscr{m}(B(w, r_2))\, \mathscr{m}(dw) - C(r_1)\, C(r_2) \qquad (3.7)\end{aligned}$$

Note that in the special case $r_1 = r_2 = r$ the above reduces to the variance of the mass in a random ball of radius r:

$$\begin{aligned}\tau(r, r) &= \operatorname{Var}(\mathscr{m}(B(W, r)))\\ &= \int \mathscr{m}(B(w, r))^2\, \mathscr{m}(dw) - C(r)^2 \qquad (3.8)\end{aligned}$$

We will prove the following:

If $W_1,\dots, W_n$ are independent and identically-distributed observations from m, the asymptotic covariance matrix U is given by $u_{ij} = 4\tau(r_i, r_j)$. Hence, the matrix V has entries given by $v_{ij} = 4\tau(r_i, r_j)/C(r_i)\, C(r_j)$.

Proof. The asymptotic covariance u_{ij} results from the fact that, in (3.1), pairs with a member in common are not independent. First consider the diagonal (variance) terms. We obtain

$$\begin{aligned}
&\operatorname{Var}\left(\binom{n}{2} C_n(r)\right)\\
&\quad = \sum_{j<k}\sum \operatorname{Var}(I_{[\|W_j\quad W_k\| \leqslant r]})\\
&\qquad + \sum_{i\neq j}\ \sum_{i\neq k}\ \sum_{j\neq k} \operatorname{Cov}(I_{[\|W_i\quad W_j\| \leqslant r]}, I_{[\|W_i\quad W_k\| \leqslant r]}) \qquad (3.9)
\end{aligned}$$

Now the first term on the rhs of (3.9) equals $\binom{n}{2}\, C(r)[1 - C(r)]$. However, the second term on the rhs of (3.9) involves $n(n-1)(n-2)$ terms which are determined by the variance of the mass in a random ball. Specifically, we calculate

$$\begin{aligned}
&\operatorname{Cov}(I_{[\|W_i\quad W_j\| \leqslant r]}, I_{[\|W_i\quad W_k\| \leqslant r]})\\
&\quad = E(I_{[\|W_i\quad W_j\| \leqslant r]} I_{[\|W_i\quad W_k\| \leqslant r]}) - E(I_{[\|W_i\quad W_j\| < r]})\, E(I_{[\|W_i\quad W_k\| < r]})\\
&\quad = \operatorname{Var}(m(B(W, r)))\\
&\quad = \tau(r, r) \qquad (3.10)
\end{aligned}$$

Hence

$$\operatorname{Var}\left(\binom{n}{2} C_n(r)\right) = \binom{n}{2} C(r)[1 - C(r)] + n(n-1)(n-2)\,\tau(r, r)$$

and therefore

$$\operatorname{Var}(n^{1/2} C_n(r)) = \frac{2C(r)[1 - C(r)]}{n-1} + \frac{4(n-2)}{n-1}\tau(r, r) \qquad (3.11)$$

This shows that $u_{ii} = 4\tau(r_i, r_i)$. The proof for u_{ij} is analogous. The form of v_{ij} follows by considering the first few terms in a Taylor expansion of $\log C_n(r)$. ▌

In the usual case that $W_1,\dots, W_n$ are correlated observations from some orbit of a dynamical system (with sufficiently good mixing properties), Theorem 1 of Denker and Keller (applied to covariances) gives

$$u_{ij} = 4\tau(r_i, r_j) + 4 \sum_{h=1}^{\infty} \kappa(h, r_i, r_j) \qquad (3.12)$$

where

$$\kappa(h, r_1, r_2) = \mathrm{Cov}(m(B(W_1, r_1)), m(B(W_{1+h}, r_2)))$$

accounts for the correlations between observations that are h iterations apart. In the presence of good mixing it is probably not necessary to estimate $\kappa(h, r_1, r_2)$ for more than a few values of h.

To obtain estimates for these quantities, first compute

$$p(j, r) = \frac{1}{n-1} \sum_{i \neq j}^{n} I_{[\|W_i \quad W_j\| \leqslant r]}$$

for each point W_j in the data set and each component r of the vector $\mathbf{r}$. Since $C_n(r) = (1/n) \sum_{j=1}^{n} p(j, r)$, the computations can be built into the algorithm for obtaining $C_n(r)$; this requires double the time of the shortest algorithm (which only counts distinct pairs), but does represent the total necessary increase in computing time required to also estimate the covariance matrix.

The estimators of $\tau(r_1, r_2)$ and $\kappa(h, r_1, r_2)$ are given by

$$\hat{\tau}(r_1, r_2) = \frac{1}{n} \sum_{j=1}^{n} p(j, r_1)\, p(j, r_2) - C_n(r_1)\, C_n(r_2)$$

$$\hat{\kappa}(h, r_1, r_2) = \frac{1}{n-h} \sum_{j=1}^{n\ \ h} p(j, r_1)\, p(j+h, r_2) - C_n(r_1)\, C_n(r_2) \qquad (3.13)$$

We applied the above technique to estimate the variability of $b(\mathbf{r})$ for the Kaplan–Yorke map at the vector $\mathbf{r} = (0.08\ 0.04\ 0.02\ 0.01\ 0.005)$ with $n = 750$. The Kaplan Yorke map in $\mathbb{R}^2$ with parameter $\lambda = 0.2$ is given by

$$T(x, y) = (2x (\mathrm{mod}\ 1), (0.2)\, y + \cos(4\pi x)) \qquad (3.14)$$

The purpose of choosing this example with these parameters and sample sizes was to compare our results to those of Denker and Keller. The details of the estimation procedure in that paper are rather sketchy, but it appears they were attempting to estimate $[\mathrm{Var}\, b(\mathbf{r})]^{1/2}$, by estimating (3.12) with four covariances $h = 1, 2, 3, 4$. Their numerical procedure either contained an error or they failed to double the values in their table, since the presented results are considerably out of line with the actual variance (knowledge about the latter being provided by the empirical variance of many estimates). To obtain a good estimate of the true variability of $b(\mathbf{r})$, we first carried out 50 independent simulations (each with sample size $n = 750$) and computed the empirical standard deviation

$$s = \left\{ \frac{1}{49} \sum_{u=1}^{50} [b_u(\mathbf{r}) - \bar{b}(\mathbf{r})]^2 \right\}^{1/2} \approx 0.0412$$

of the individual slopes $b_1(\mathbf{r}),\dots, b_{50}(\mathbf{r})$ from their arithmetic mean $\bar{b}(\mathbf{r})$. This value compares favorably with the empirical standard deviation (~ 0.035) of the 20 slopes provided by Denker and Keller, but is twice the size of their estimates of the standard deviation. We then ran three independent simulations and obtained estimates of (3.12) [using (3.13)] first ignoring all covariance terms $\kappa(h, r_i, r_j)$ and then including the first four terms $h = 1$, 2, 3, 4. Our estimates of the standard deviation of $b(\mathbf{r})$ [obtained by estimating V from (3.13) and then substituting into the second line of (3.6)] are given below (the first component in each pair corresponds to the estimate with no covariances included):

$$(0.0380, 0.0434) \qquad (0.0359, 0.0396) \qquad (0.0404, 0.0496)$$

This appears to be an extremely effective as well as simple procedure for obtaining accurate standard errors, and it is to be hoped that this technique would be applied routinely when estimating the correlation dimension from least squares.

We close this section with some general remarks on estimation. It is clear from Section 3.1 that we should only be estimating ν at small values of r. However, we can expect greater variability and poorer estimates when r is smaller. The relative rates at which $r \to 0$ and $n \to \infty$ become important here (and the optimal rates are likely to depend on the true underlying dimension and certain specific scaling properties of m). Various authors have considered the problem of determining the amount of data necessary in order to make valid inferences about dimension. Nicolis and Nicolis,(42) Grassberger,(29) Essex *et al.*,(26) and Essex(25) is a particularly interesting sequence of papers which, in part, deal with this question. An important conclusion to emerge from these papers is that most "real" data sets are not nearly large enough to provide good inference about dimension. However, we point out that the efforts toward determining the sample size necessary to observe a "scaling region" in r (see in particular Essex *et al.*(26) and Essex(25)) may only really be applicable to measures which scale on levels very much like D-dimensional Lebesgue measure, since this is the distribution for which computations were carried out. (Since Lebesgue measure "scales" well at all reasonable values of r, the problem of sample size here reduces to guaranteeing the existence of some r range over which the data set is neither saturated nor depleted.) In fact, the typical search for a scaling region in r often implicitly ignores the asymptotic ($r \to 0$) nature of ν; larger values of r frequently are eliminated only when "saturation" or "turning" is observed. For distributions which scale correctly only for small values of r the sample sizes necessary to make valid inferences may be even greater than the (already large) estimates provided by Essex.(25) We also note that the "boundary corrections" proposed by Essex *et al.*[26] (to increase the range of the scaling region) should be employed with great caution; since ν is the scaling exponent of an average ball, computation of ν does in fact require us to average over all points, including boundary points. When we focus on the scaling behavior of individual points in the data set we are really looking at σ. Thus, making too many boundary corrections may result in a statistic which is estimating neither ν nor σ, but something in between.

See also Ramsey and Yuan[48] for a discussion of estimation problems and sample size.

4. ESTIMATING THE INFORMATION DIMENSION

Throughout this section we will assume that

$$\lim_{r \to 0} \frac{\log m(B(x, r))}{\log r} = \sigma \quad m\text{-a.s.}$$

Under this assumption the least squares techniques discussed in the preceding section can also be employed here with $m(B(x, r))$ in the role of $C(r)$ (provided x is a generic point on the attractor). That is, we consider the identity

$$\log m(B(x, r)) = \alpha(x, r) + \sigma \log r \tag{4.1}$$

The sample proportion

$$p_n(r) = \frac{1}{n} \sum_{j=1}^{n} I_{[\|W_j - x\| \leqslant r]}$$

provides a point estimate of $m(B(x, r))$. The difficulties with least squares estimation noted in Section 3 will of course also occur here, and in fact we can expect even greater unpredictability in the intercept behavior because of local effects at the point x.

Cutler and Dawson[19,20] studied the dimension-related properties of data from a large class of distributions. While the results in the second paper are formulated in terms of distributions on the unit interval, it is clear how to extend them to a more general setting in higher dimensions. We provide the basic constructions now.

Let $\mathscr{M}$ be a cube in $\mathbb{R}^N$; without loss of generality we can take $\mathscr{M}$ to be the unit cube. Let $r \geqslant 2$ be a fixed positive integer, and let $\mathbf{p} = (p_1, p_2, ..., p_{r^N})$ be a vector of probabilities, where $\sum_{j=1}^{r^N} p_j = 1$. We allow the possibility $p_j = 0$ for some j, but at least two should be nonzero to avoid trivial constructions. Now, for each $n \geqslant 1$, $\mathcal{M}$ can be split uniquely into r^{nN} nonoverlapping cubes of side length r^{-n}. At the first stage ($n = 1$) label the r^N cubes as $c_1, ..., c_{r^N}$ (maintaining this pattern of labeling at further stages, so that the correspondence between the second-stage subcubes of c_j and the labels $c_{j,1}, c_{j,2}, ..., c_{j,r^N}$ is fixed and determined by the first-stage pattern). To the nth-stage subcube $c_{j_1 \cdots j_n}$ we assign the probability $m_*(c_{j_1 \cdots j_n}) = p_{j_1} p_{j_2} \cdots p_{j_n}$. Thus, the stages are taken to be independent. (More general probability structures are possible, but for simplicity we will consider only the independent model here.) The measure m_*, having been defined over r-adic cubes, now extends uniquely to a probability measure on $\mathcal{M}$. Letting g be any nonnegative function such that $\int_{\mathcal{M}} g(x)\, m_*(dx) = 1$, we now define m by $m(B) = \int_B g(x)\, m_*(dx)$. In this context $g = dm/dm_*$ is the density (or Radon–Nikodym derivative) of m with respect to the canonical measure m_*.

The measure m falls into one of three categories, which we call *smooth*, *semifractal*, and *fractal*. The measure m is smooth if $p_j = r^{-N}$ for each j (so m_* coincides with Lebesgue measure on $\mathcal{M}$). Hence, smooth measures are those which have a density with respect to Lebesgue measure. (See Examples 4.1, 4.2, and 4.6 in this paper.) A more general definition of smooth is possible by considering measures which have densities over smooth manifolds (this approach is taken in Cutler and Dawson,[19] where such measures are called "D-regular"). Semifractal measures (so-called because they have properties in common with both smooth and fractal measures) occur when at least one $p_j = 0$ and all nonzero p_j take on a common value. (That is, each nonzero probability $p_j = 1/k$, where k, $2 \leqslant k \leqslant r^N - 1$, is the total number of nonzero p_j.) As a result, m_* is the uniform measure on a Cantor set (where the Cantor set is obtained by eliminating all empty, i.e., zero-probability, subcubes from $\mathcal{M}$), and m has a density with respect to this uniform measure. (See Examples 4.3 and 4.8.) Finally, the fractal measures occur when there are at least two nonzero p_j which are not equal. The closed support of a fractal measure may be $\mathcal{M}$ (if each $p_j > 0$) or a Cantor subset of $\mathcal{M}$ (if at least one $p_j = 0$). However, a fractal measure m is always singular with respect to the uniform measure on its closed support, and this distinguishes the fractal case from the first two. (See Examples 4.4, 4.5, and 4.9.)

The intercept $\alpha(x, r) = \log[m(B(x, r))/r^\sigma]$ exhibits three distinct behaviors (as $r \to 0$), depending on whether m is smooth, semifractal, or fractal. Let g denote the density function dm/dm_* defined above. In the smooth case, $\alpha(x, r)$ converges to $\log[K_\sigma g(x)]$; K_σ is a normalizing constant which depends only on σ. (In our simple cube construction in $\mathbb{R}^N$, we always have $\sigma = N$ in the smooth case.) In the semifractal case, $\alpha(x, r)$ asymptotically oscillates between two fixed bounds $a_1 + \log g(x)$ and

$a_2 + \log g(x)$, where a_1 and a_2 are determined by the geometry of the supporting Cantor set. The fractal case exhibits the worst behavior; for m-almost all points x, we observe that

$$\liminf_{r \to 0} \alpha(x, r) = -\infty \qquad \text{and} \qquad \limsup_{r \to 0} \alpha(x, r) = +\infty$$

Some order is restored in this last case by the fact that when averaging over random points X selected from m we can show that $\alpha(X, r)/(\log 1/r)^{1/2}$ tends to a mean-zero Gaussian distribution as $r \to 0$. Hence $|\alpha(X, r)| = o((\log 1/r)^{1/2+\varepsilon})$.

Guckenheimer[34] suggested sorting the distances $d_j = \| W_j - x \|$ into ascending order, then plotting the logarithms of the ordered distances against the fixed values $\log(j/n)$, $j = 1, 2,..., n$. (In this approach the radii become random and the observed proportions fixed, which is the reverse of the earlier procedure.) The appealing feature of this method is that it is not necessary to select a vector $\mathbf{r}$ and the radii at the lower end of the scale automatically decrease toward 0 as $n \to \infty$. However, we would need a rule for removing observations at the upper end of the scale (large radii), and the covariance structure of the logarithms of the ordered distances would also have to be determined (as in Section 3.2) in order to be able to estimate the variability of the slope of the least squares line. It is also important to note that reversing the role of the radii and the proportions means that the slope of the least squares line is inverted, leading naturally to an estimate of $1/\sigma$ rather than σ. Taking the reciprocal to obtain a point estimate of σ does not always produce satisfactory results, and so it is particularly important here to have a good sense of the error bounds on estimates. Termonia and Alexandrowicz[59] seem to have suggested a somewhat similar procedure; furthermore, they average over several points x to reduce the influence of local effects. However, since this averaging is performed prior to scaling (i.e., before taking logarithms), we would expect their procedure to perform poorly in the case of measures with densities with sharp singularities (perhaps leading to estimates of v rather than σ.)

Our approach to estimating the information dimension will be that of using a nearest neighbor statistic (the extreme low end of the scale in Guckenheimer's method) to construct confidence intervals for σ. An interval $[c_l, c_u]$ is called a 95% *confidence interval* for σ if the statistical procedure used to construct $[c_l, c_u]$ produces intervals which, in the long run, cover σ 95 times out of 100. Our emphasis is on intervals rather than point estimates for three reasons. First, a central problem in dimension estimation has been the difficulty in obtaining accurate error bounds; this problem is directly addressed by confidence intervals. Second, it is likely that in most situations accurate intervals for dimension will be all that is really needed. Finally, the point estimate produced by our method is not particularly well behaved (being a reciprocal), so in using this procedure it is best to regard the point estimate with skepticism and look instead to the confidence interval. This will be discussed again later.

The result which forms the basis for our procedure is the following: If $W_1, W_2, \ldots$ is a sequence of independent observations from $\mathscr{m}$ and $d_n(x) = \min_{1 \leqslant j \leqslant n} \|W_j - x\|$, then, provided the dimension regularity condition $\sigma_H(x) = \sigma_P(x) = \sigma$ holds for $\mathscr{m}$-almost all x, we have

$$\lim_{n \to \infty} \frac{\log d_n(x)}{\log 1/n} = \frac{1}{\sigma} \qquad \text{with probability 1} \quad \mathscr{m}\text{-a.s.}$$

The convergence of this sequence, however, is usually very slow, with the local density $g(x)$ at x playing a significant role. This is not surprising, since the distribution of points near x is affected by the density value at x. A high (low) value of $g(x)$ can make the local dimension at x appear lower (higher) than it really is, particularly in smaller samples. Define the linear transformation $L_n(x)$ by

$$L_n(x) = \sigma \log n \left(\frac{\log d_n(x)}{\log 1/n} - \frac{1}{\sigma} \right) \tag{4.2}$$

Cutler and Dawson[19] show that if $\mathscr{m}$ is a smooth measure, then for $\mathscr{m}$-almost all x, the distribution of $L_n(x)$ converges, as $n \to \infty$, to an *extreme value* distribution (see Appendix C for details) with mean $\gamma + \log[K_\sigma g(x)]$ (where $\gamma =$ Euler's constant ≈ 0.577). Thus, the center of the limiting distribution is determined by $g(x)$. Note also that this result illustrates the poor convergence behavior of the difference

$$\frac{\log d_n(x)}{\log 1/n} - \frac{1}{\sigma}$$

Inflation by $\log n$ is sufficient to produce a nondegenerate error distribution. Furthermore, this turns out to be the best possible scenario; in the case of fractal measures the behavior of $L_n(x)$ is so erratic that its distribution does not converge as $n \to \infty$. Averaging over points x in the space as well as damping by the quantity $(\log n)^{1/2}$ must be carried out to obtain a limiting distribution. The quantity $L_n(X)/(\log n)^{1/2}$ approaches a mean-zero Gaussian distribution as $n \to \infty$. (See Appendix C for more details.) Although the density function g does not appear explicitly in the asymptotic distribution in the fractal case, this is due to the $(\log n)^{1/2}$ divisor, which slowly drives the mean toward zero. The rate of convergence to the mean-zero Gaussian will be affected by the nature of g.

One way of eliminating the local density effect at x is to obtain two independent random samples $W_{1,1},\ldots, W_{1,n}$ and $W_{2,1},\ldots, W_{2,mn}$, one of size n and the other of size mn (where $m \geqslant 2$ is some fixed real number). Compute the respective nearest-neighbor distances

$$d_{1,n}(x) = \min_{1 \leqslant j \leqslant n} \|W_{1,j} - x\| \qquad \text{and} \qquad d_{2,mn}(x) = \min_{1 \leqslant j \leqslant mn} \|W_{2,j} - x\|$$

and define the log-ratio statistic

$$R_n(x) = \frac{1}{\log m} \log\left(\frac{d_{1,n}(x)}{d_{2,mn}(x)}\right) \tag{4.3}$$

It can be shown (see Appendix C) that $R_n(x)$ is asymptotically free of the local density effect. We now present a procedure, based on this statistic, which can be applied to smooth, semifractal, and fractal measures. The exact range of measures for which this method is valid is not known, but the asymptotic theory can be proven rigorously in certain cases and validated numerically in many others. We will present the basic result and procedure first, then discuss the general method of proof. Rigorous details can be found in Appendices B and C.

Let $S_{1,n} = \{W_{1,1},\ldots, W_{1,n}\}$ and $S_{2,mn} = \{W_{2,1},\ldots, W_{2,mn}\}$ denote two independent random samples from $\mathscr{m}$. (Of course in practice the observations will not be independent, and here our mixing assumptions become important. The practical details of how to sample will be discussed later.) Let $S_n = S_{1,n} \cup S_{2,mn}$ and then select, independently of S_n, a random sample $B = \{X_1,\ldots, X_{k_n}\}$ from $\mathscr{m}$. We refer to B as the set of *basepoints*, and use the notation X_j to distinguish the basepoints from the data points W_j (although all are observations from $\mathscr{m}$). The number k_n of basepoints will generally be much smaller than n, and determining the correct ratio of k_n to n is still an open problem, although we can offer some practical guidelines. For each basepoint compute the ratio statistic $R_n(X_j)$ defined in (4.3), the empirical mean $\bar{R}_n = (1/k_n) \sum_{j=1}^{k_n} R_n(X_j)$, and the empirical standard deviation

$$s = \left\{\frac{1}{k_n - 1} \sum_{j=1}^{k_n} [R_n(X_j) - \bar{R}_n]^2\right\}^{1/2}$$

Let $\theta(S_n)$ denote the conditional variance of $R_n(X_j)$ given the combined sample S_n. That is, $\theta(S_n) = E((R_n(X_j) - \mu(S_n))^2 \mid S_n)$, where $\mu(S_n) = E(R_n(X_j) \mid S_n)$ is the conditional mean of $R_n(X_j)$ given S_n. (It follows then that s is a point estimate of $[\theta(S_n)]^{1/2}$.) We propose the following:

Theorem 4.1. Let

$$Z_{n,k_n}=\frac{\sum_{j=1}^{k_n} R_n(X_j)-k_n/\sigma}{[k_n\theta(S_n)]^{1/2}}$$

Then, provided $n \to \infty$, $k_n \to \infty$, and $k_n/n \to 0$ at a suitable rate, the distribution of Z_{n,k_n} approaches a standard Gaussian distribution with mean zero and variance 1. An approximate 95% confidence interval for σ is given by $[c_l, c_u]$, where $c_l=(\bar{R}_n+2s/\sqrt{k_n})^{-1}$ and $c_u=(\bar{R}_n-2s/\sqrt{k_n})^{-1}$ and s is the empirical standard deviation of the $R_n(X_j)$.

Several ideal lie behind the above central limit theorem. The three key ideas are the following:

1. For large n, the variables $R_n(X_j)$ center around the value $1/\sigma$. This centering should occur fairly quickly, as the bias due to a density effect has been removed. The variance of $R_n(X_j)$ (as a function of n) stays bounded for smooth and semifractal measures, actually converging in the smooth case to the asymptotic value $\theta(\sigma, m)=\pi^2/[3(\log m)^2 \sigma^2]$. It is instructive to compare this value with the empirical variation obtained in the smooth Examples 4.1, 4.2, and 4.6. For fractal measures the variance grows at a rate proportional to $\log n$, the constant of proportionality being larger when $\mathcal{M}$ is "more fractal." However, the growth in variability is sufficiently controlled that central limit theory is still applicable. See Appendix C.

2. For large n and k_n with k_n/n in the appropriate ratio, the conditional variance $\theta(S_n)$ gives a good approximation to the true variability of the numerator (this is important because the conditional variance is the quantity naturally estimated from the data.) This is a consequence of Theorems B.1 and B.2 in Appendix B.

3. The variables $R_n(X_j)$, $n=1, 2,...,$ $j=1,..., k_n$, form an array in which each row $R_n(X_1),..., R_n(X_{k_n})$ can be embedded in an infinite exchangeable sequence. (A sequence of random variables is called *exchangeable* if any permutation of any k of them has the same joint distribution as the first k variables.) This embedding results by considering an infinite sequence of random basepoints $X_1, X_2,...$ and regarding $R_n(X_1),..., R_n(X_{k_n})$ as the first k_n terms of the corresponding infinite sequence $R_n(X_1)$, $R_n(X_2)$,... . Exchangeability is a property related to but weaker than i.i.d. (independent and identically-distributed). The variables $R_n(X_1),..., R_n(X_{k_n})$ fail to be independent because each is a function of the same sample S_n. However, as we move down the array, the sample size n increases and we expect the nearest-neighbor distance to one basepoint to become independent of the nearest-neighbor distance to any other basepoint. Bickel and Breiman[5] have developed some related limit

theorems involving functions of nearest neighbors; however, they use the observations simultaneously as basepoints (which destroys the ability to embed into an infinite exchangeable sequence) and their method of normalization does not allow covariances to vanish quickly enough. Techniques for proving limit theorems for infinite-exchangeable arrays are known, and in Appendix B we use a generalization of a method suggested by problems in Chow and Teicher (ref. 12, #3, 4, p. 317).

Assuming that Z_{n,k_n} is approximately Gaussian with mean 0 and variance 1, we conclude that there is a 95% probability of Z_{n,k_n} falling between ± 2. This translates into a 95% probability of the random interval $\bar{R}_n \pm 2[\theta(S_n)/k_n]^{1/2}$ covering $1/\sigma$. Inverting this, we obtain a confidence interval for σ:

$$\left[\frac{1}{\bar{R}_n + 2[\theta(S_n)/k_n]^{1/2}}, \frac{1}{\bar{R}_n - 2[\theta(S_n)/k_n]^{1/2}}\right]$$

At this stage the unknown quantity $[\theta(S_n)]^{1/2}$ is replaced by its estimate s, the empirical standard deviation. This results in the approximate 95% confidence interval $[c_l, c_u]$ presented in Theorem 4.1.

The empirical mean $\bar{R}_n$ will be approximately distributed as a Gaussian for large n, and provides a point estimate of $1/\sigma$. The reciprocal $\hat{\sigma} = 1/\bar{R}_n$ can be used as a point estimate of σ; however, as the reciprocal of a Gaussian, $\hat{\sigma}$ has no moments. Numerically this is observed as wild swings in the value of $\hat{\sigma}$ from simulation to simulation. For this reason point estimation of σ is not particularly good with this method. We also note for the interval $[c_l, c_u]$ proposed in Theorem 4.1 that the increase in coverage probability toward 95% grows very slowly as a function of n. We found that a much more efficient way to achieve 95% coverage was to widen the confidence interval; for example, replacing "2" by "3" in the definition of $[c_l, c_u]$ generally produced 95% or better coverage often with half the sample size required by the first method. This is an important consideration when running short on data. The price is slightly wider confidence intervals.

We now present a series of examples with tables illustrating the above technique. There are three sample size quantities involved; k = number of basepoints, n = size of first sample, and m = multiplicative factor determining the size of the second sample. The total effective sample size (not including discarded iterates) is $t = k + n + mn$. In each case 20 typical confidence intervals with corresponding point estimates $\hat{\sigma}$ (obtained from repeated independent simulations) are provided, including one interval which fails to cover σ (marked with an asterisk $*$). The ratio 1 in 20 reflects the approximate probability of failing to cover. Observed coverage probabilities (based on 50 simulations unless otherwise indicated) are given in the tables. The purpose is to provide the reader with a feeling for the behavior of the intervals and point estimates produced by this method. The table columns marked s, $R_{\min}$, and $R_{\max}$, provide, respectively, the empirical standard deviation s, the observed minimum, and the observed maximum, of the $R_n(X_j)$. The purpose is to contrast the behavior of the $R_n(X_j)$ in the smooth and nonsmooth cases. In fact it is often possible to distinguish fractal measures from smooth measures simply by observing s, rmin, and rmax. This will be indicated at the appropriate places.

The following five examples are of dynamical systems on the unit interval. In order to facilitate comparison among these examples, the same sample size controls $k = 300$, $n = 500$, and $m = 5.0$ were used in each case. This produced a total effective sample size of $t = 3300$, which is quite modest by comparison with the sample sizes often used with other methods. To obtain shorter confidence intervals, k and n should be increased with $k/n \to 0$. Certain numbers of iterates often must be discarded. We found that the crucial assumptions to be satisfied were independence of the two samples and independence among the basepoints.

Table III. Example 4.1[a]

$\hat{\sigma}$	$[c_l, c_u]$	s	$R_{\min}$	$R_{\max}$
1.008	[0.887, 1.166]	1.167	−2.212	4.119
1.052	[0.930, 1.211]	1.081	−4.340	3.974
1.001	[0.893, 1.140]	1.052	−2.938	4.947
0.976	[0.858, 1.131]	1.218	−3.543	5.607
0.985	[0.862, 1.148]	1.252	−4.128	6.978
0.952	[0.855, 1.074]	1.035	−2.837	4.967
1.105	[0.970, 1.284]	1.094	−4.082	3.933
1.021	[0.898, 1.183]	1.162	−2.523	6.129
0.935	[0.834, 1.063]	1.117	−3.056	4.484
0.932	[0.832, 1.060]	1.119	−3.264	4.992
0.872	[0.783, 0.984]*	1.129	−1.450	5.280
0.969	[0.863, 1.107]	1.109	−2.320	5.424
1.114	[0.979, 1.292]	1.073	−2.400	4.446
1.091	[0.955, 1.273]	1.132	−3.482	6.011
1.063	[0.928, 1.244]	1.186	−2.856	5.071
0.956	[0.849, 1.093]	1.138	−5.917	4.223
1.071	[0.937, 1.251]	1.162	−4.671	4.388
0.928	[0.829, 1.053]	1.110	−2.319	5.056
1.052	[0.923, 1.221]	1.144	−3.131	5.240
0.977	[0.871, 1.112]	1.076	−1.721	4.908

[a] $\sigma = 1$, $k = 300$, $n = 500$, $m = 5.0$, $t = 3300$, observed coverage ≈94%.

(This is generally achieved in the presence of mixing, for example, by separating basepoints by many iterates. The intermediate iterates are discarded.) Independence of the observations within each sample was found to be less critical; except in cases where mixing was very slow, we were able to select successive observations when constructing a sample.

Example 4.1. The Logistic Map. Here we consider $T(x) = 4x(1-x)$ as in Example 3.2. Refer to Table III. Simulations were performed by generating an initial condition at random from $[0, 1]$ and then iterating under T. The first 100 iterates were discarded. The jth and $(j+1)$th basepoints were separated by $3j$ iterates. (However, a fixed separation of 10–15 iterates probably would have been sufficient.) The two samples were separated by 50 iterates. Observations within each of the two samples were selected successively. Compare the observed values of s with the asymptotic value $[\theta(1, 5)]^{1/2} \approx 1.127$ for smooth measures with $\sigma = 1$ and $m = 5$.

Example 4.2. Here $T(x)$ is the map discussed in Example 3.3 with $\gamma = 1/3$. As in the previous example, an initial condition was selected randomly from $[0, 1]$, the first 100 iterates discarded, and the basepoints

Table IV. Example 4.2[a]

$\hat{\sigma}$	$[c_l, c_u]$	s	$R_{\min}$	$R_{\max}$
0.981	[0.876, 1.113]	1.049	−1.592	4.148
1.263	[1.085, 1.512]*	1.128	−2.996	4.347
0.972	[0.867, 1.104]	1.072	−3.357	4.096
1.029	[0.907, 1.190]	1.136	−3.178	4.250
0.966	[0.854, 1.086]	1.172	−3.171	4.124
0.944	[0.836, 1.086]	1.195	−2.691	5.239
1.062	[0.924, 1.200]	1.135	−2.039	4.548
1.083	[0.956, 1.249]	1.065	−3.941	4.403
1.110	[0.979, 1.280]	1.040	−3.888	3.538
1.043	[0.918, 1.207]	1.132	−3.757	3.956
0.925	[0.828, 1.049]	1.105	−1.854	4.441
0.925	[0.827, 1.049]	1.110	−2.611	4.378
1.112	[0.978, 1.289]	1.068	−3.290	4.452
0.943	[0.836, 1.081]	1.171	−2.543	6.295
1.054	[0.927, 1.222]	1.129	−2.545	4.408
0.951	[0.842, 1.093]	1.180	−2.847	5.458
1.092	[0.962, 1.262]	1.071	−2.561	3.613
0.974	[0.867, 1.111]	1.094	−2.994	5.028
1.044	[0.915, 1.215]	1.167	−3.616	6.439
1.024	[0.902, 1.286]	1.147	−2.774	4.624

[a] $\sigma = 1$, $k = 300$, $n = 500$, $m = 5.0$, $t = 3300$, observed coverage ≈95%.

sampled $3j$ iterates apart (in this example such separation is probably necessary because of slow mixing). We found that using successive observations in the samples produced poor coverage probabilities (approximately 84% instead of the intended 95%). This problem was rectified by discarding j iterates between the $(j-1)$th and jth observations to improve the mixing. Refer to Table IV. Again compare the observed values of s with 1.127.

Example 4.3. The Cantor Distribution. This corresponds to Example 3.1 and the uniform distribution m_K on the Cantor set. This is a semifractal measure. The asymptotic behavior of nearest neighbors for semifractal measures has more in common with smooth measures than fractal measures. In particular, $\theta(S_n)$ stays bounded as $n \to \infty$. Note, however, the increase in the observed values of s as compared to the two preceding examples. (See Table V.) Successive observations were used within each sample; basepoints were sampled 50 iterates apart. Iteration was performed by shifting a binary string of 50 0's and 2's once to the left and randomly generating a new 0 or 2 (with equal probability) into the

Table V. Example 4.3[a]

$\hat{\sigma}$	$[c_l, c_u]$	s	$R_{\min}$	$R_{\max}$
0.580	[0.516, 0.661]	1.838	−2.763	7.619
0.765	[0.666, 0.900]*	1.691	−5.930	8.391
0.585	[0.519, 0.670]	1.885	−4.609	7.501
0.673	[0.599, 0.768]	1.597	−5.407	5.969
0.587	[0.521, 0.673]	1.881	−4.663	8.121
0.598	[0.538, 0.673]	1.622	−2.769	6.877
0.626	[0.551, 0.723]	1.872	−4.857	9.227
0.679	[0.594, 0.792]	1.820	−4.479	7.648
0.702	[0.617, 0.815]	1.712	−5.520	6.111
0.719	[0.616, 0.863]	2.014	−7.506	9.609
0.616	[0.547, 0.705]	1.771	−4.065	8.043
0.656	[0.574, 0.767]	1.900	−3.405	7.894
0.610	[0.545, 0.692]	1.691	−3.989	6.556
0.598	[0.530, 0.687]	1.872	−7.812	10.176
0.576	[0.517, 0.650]	1.719	−3.423	6.537
0.647	[0.570, 0.748]	1.813	−3.345	6.839
0.617	[0.550, 0.704]	1.721	−3.066	7.576
0.657	[0.578, 0.760]	1.791	−3.877	7.115
0.644	[0.565, 0.749]	1.884	−4.163	7.645
0.580	[0.516, 0.662]	1.847	−3.334	11.825

[a] $\sigma = 0.6309$, $k = 300$, $n = 500$, $m = 5.0$, $t = 3300$, observed coverage ≈94%.

50th position. (The string of 0's and 2's is interpreted as the ternary expansion of the corresponding point x.)

Example 4.4. A Fractal Cantor Distribution. We consider the ternary map $T(x) = 3x \pmod 1$ as in the preceding example, but 0's and 2's are generated in unequal proportions. The resulting distribution is supported on the Cantor set K, but is singular with respect to the uniform distribution m_K on that set. We chose $p = 0.2$, where p is the probability of generating a "2" into the binary string. It follows that

$$\sigma = -\frac{0.8 \log 0.8 + 0.2 \log 0.2}{\log 3} \approx 0.4555$$

Note the sharp increase in the size of s and the range of the $R_n(X_j)$ in Table VI. The size of p represents a limitation on this method; as $p \to 0$, the fluctuations of $R_n(X_j)$ increase, and a corresponding increase in both k and n is required to maintain coverage probabilities.

Example 4.5. A Fractal Binary Distribution. Here we take the binary map $T(x) = 2x \pmod 1$. Iteration is performed by shifting a

Table VI. Example 4.4[a]

$\hat{\sigma}$	$[c_l, c_u]$	s	$R_{\min}$	$R_{\max}$
0.465	[0.402, 0.552]	2.932	−6.252	13.852
0.431	[0.375, 0.507]	3.002	−6.484	15.675
0.500	[0.430, 0.596]	2.800	−9.265	11.632
0.425	[0.374, 0.491]	2.753	−7.760	14.365
0.519	[0.440, 0.633]	3.003	−8.181	13.983
0.498	[0.430, 0.591]	2.742	−6.167	11.694
0.403	[0.358, 0.461]	2.704	−3.736	13.726
0.410	[0.357, 0.481]	3.125	−9.128	15.767
0.429	[0.377, 0.499]	2.798	−4.160	14.503
0.583	[0.493, 0.713]*	2.713	−6.395	13.914
0.486	[0.418, 0.580]	2.899	−8.947	15.675
0.465	[0.398, 0.558]	3.104	−7.749	15.692
0.491	[0.426, 0.580]	2.704	−9.645	11.368
0.420	[0.368, 0.488]	2.889	−6.448	15.182
0.407	[0.359, 0.470]	2.857	−5.899	11.537
0.491	[0.422, 0.586]	2.856	−7.331	11.275
0.427	[0.373, 0.499]	2.935	−11.572	16.969
0.416	[0.364, 0.484]	2.956	−4.787	17.638
0.439	[0.384, 0.513]	2.849	−6.605	11.868
0.448	[0.391, 0.524]	2.797	−6.203	14.783

[a] $\sigma = 0.4555$, $k = 300$, $n = 500$, $m = 5.0$, $t = 3300$, observed coverage ≈92%.

binary string of 50 0's and 1's once to the left and generating a new 0 or 1 into the 50th position. (This method of iteration is necessary in this particular example because iterating this map algebraically results in all initial conditions being attracted to 0 within a finite number of steps. It is probably not necessary for the ternary map in Examples 4.3 and 4.4, but it does nicely illustrate the nature of the mixing property of these shift maps.) By generating 0's and 1's in unequal proportions, the resulting measure is singular with respect to Lebesgue measure. We set $p = 0.2$, where p is the long-run proportion of 1's. Hence, this is an example where the attractor, being [0, 1], is not a fractal set, but the distribution in question is a fractal measure, having

$$\sigma = -\frac{0.8 \log 0.8 + 0.2 \log 0.2}{\log 2} \approx 0.7219$$

Refer to Table VII.

The following are examples of dynamical systems with higher-dimensional phase spaces. Because of limits on the computing equipment

Table VII. Example 4.5[a]

$\hat{\sigma}$	$[c_l, c_u]$	s	$R_{\min}$	$R_{\max}$
0.746	[0.651, 0.872]	1.679	−5.594	7.952
0.735	[0.640, 0.863]	1.746	−4.247	9.129
0.725	[0.622, 0.869]	1.982	−5.530	11.060
0.640	[0.562, 0.742]	1.871	−3.900	8.081
0.737	[0.646, 0.858]	1.656	−6.785	9.296
0.634	[0.563, 0.725]	1.711	−3.632	7.966
0.642	[0.572, 0.731]	1.637	−2.118	8.196
0.695	[0.608, 0.811]	1.781	−4.766	7.745
0.776	[0.677, 0.909]	1.630	−3.353	8.001
0.763	[0.668, 0.890]	1.613	−5.572	7.152
0.644	[0.561, 0.755]	1.986	−6.328	8.986
0.775	[0.677, 0.906]	1.623	−4.618	6.885
0.704	[0.626, 0.804]	1.524	−3.249	6.017
0.709	[0.623, 0.823]	1.687	−5.756	6.501
0.777	[0.678, 0.908]	1.615	−3.163	8.943
0.679	[0.592, 0.797]	1.876	−5.145	9.215
0.739	[0.640, 0.873]	1.802	−3.654	9.085
0.614	[0.541, 0.708]*	1.890	−4.272	8.061
0.703	[0.624, 0.805]	1.556	−5.476	6.632
0.708	[0.622, 0.821]	1.689	−5.053	6.640

[a] $\sigma = 0.7219$, $k = 300$, $n = 500$, $m = 5.0$, $t = 3300$, observed coverage ≈95%.

available and the prohibitive computing time required to do many simulations on small machines, it was not possible to obtain examples of systems with a large information dimension. The largest system studied was that of uniform random vectors in the unit cube in $\mathbb{R}^5$.

Example 4.6. Uniform Distribution in $\mathbb{R}^5$. Five-dimensional vectors were generated in the unit cube using a pseudorandom number generator. All basepoints and observations were sampled successively because of the good mixing properties of the generator. For $k = 300$ and $m = 5$, we found $n = 4000$ to be the minimum sample size able to provide approximate 95 % coverage. However, by using wider intervals (replace "2" by "3" in $[c_l, c_u]$) we found coverage in excess of 95 % with $n = 3000$. Note from Table VIII that the point estimate $\hat{\sigma}$ usually underestimates the true value $\sigma = 5$; this occurs because, for $n = 4000$, the expected value of $R_n(X_j)$ slightly exceeds 1/5. This bias trails off very slowly as n increases. Compare the observed values of s with $[\theta(5, 5)]^{1/2} \approx 0.2254$, the asymptotic standard deviation for smooth measures with $\sigma = 5$ and $m = 5$.

In the next three examples we used $k = 400$ and $m = 7.0$. This was done

Table VIII. Example 4.6[a]

$\hat{\sigma}$	$[c_l, c_u]$	s	$R_{\min}$	$R_{\max}$
4.787	[4.269, 5.447]	0.219	−0.563	1.044
4.661	[4.117, 5.369]	0.245	−0.565	1.154
5.164	[4.537, 5.991]	0.232	−0.494	0.981
4.878	[4.308, 5.621]	0.235	−0.471	1.070
4.193	[3.762, 4.736]*	0.237	−0.528	1.113
4.720	[4.201, 5.387]	0.227	−0.533	0.964
4.720	[4.179, 5.422]	0.238	−0.498	1.124
4.961	[4.364, 5.746]	0.239	−0.814	0.942
4.970	[4.385, 5.734]	0.232	−0.723	0.781
5.084	[4.495, 5.849]	0.223	−0.614	0.930
4.533	[4.076, 5.104]	0.214	−0.426	0.917
5.296	[4.607, 6.226]	0.244	−0.901	0.894
4.799	[4.228, 5.547]	0.244	−0.649	1.023
5.081	[4.434, 5.948]	0.249	−0.919	0.820
4.660	[4.090, 5.414]	0.259	−0.753	1.276
4.834	[4.282, 5.549]	0.231	−0.684	1.011
4.682	[4.166, 5.345]	0.229	−0.498	0.915
5.038	[4.401, 5.889]	0.249	−1.099	0.900
4.900	[4.321, 5.654]	0.236	−0.615	0.817
4.552	[4.056, 5.187]	0.233	−0.494	1.220

[a] $\sigma = 5$, $k = 300$, $n = 4000$, $m = 5.0$, $t = 24300$, observed coverage ≈95 %.

to shorten the confidence intervals; k appears in the width of $[c_l, c_u]$ and larger values of m have a damping effect on variance. (However, it is not clear that increasing m is of sufficient value to justify the corresponding large increases in total sample size.)

Example 4.7. The Hénon Mapping. Here we consider the well-known mapping[36] in $\mathbb{R}^2$: $T(x, y) = (y + 1 - ax^2, bx)$ with the standard parameter values $a = 1.4$ and $b = 0.3$. It is known that $1.21 \leqslant \sigma \leqslant 1.25$, and evidence suggests that the corresponding distribution is semifractal, the attractor being Cantorian in one direction and composed of smooth lines in the other. The first 1000 iterates from this system were discarded, basepoints were sampled 20 iterates apart (this was found to be critical; sampling close together gave very poor results), and observations were sampled ten iterates apart (this may not have been necessary). See Table IX. For smooth measures in $\mathbb{R}^2$ with $m = 7.0$, we have $[\theta(2, 7)]^{1/2} \approx 0.4661$.

Example 4.8. Cantor Distribution in $\mathbb{R}^3$. Here we consider the Cartesian product $K \times K \times K$ of the standard Cantor set K in $[0, 1]$, and

Table IX. Example 4.7[a]

$\hat{\sigma}$	$[c_l, c_u]$	s	$R_{\min}$	$R_{\max}$
1.275	[1.165, 1.408]	0.740	−1.529	3.932
1.241	[1.135, 1.368]	0.749	−2.248	3.615
1.231	[1.123, 1.363]	0.784	−1.378	4.549
1.128	[1.036, 1.238]*	0.784	−2.331	3.670
1.184	[1.088, 1.298]	0.745	−1.494	3.490
1.229	[1.131, 1.345]	0.704	−1.438	3.060
1.211	[1.112, 1.330]	0.738	−1.418	4.303
1.247	[1.137, 1.380]	0.776	−1.680	3.932
1.272	[1.155, 1.417]	0.802	−2.555	3.549
1.190	[1.085, 1.318]	0.814	−1.853	4.125
1.189	[1.096, 1.298]	0.713	−1.844	4.115
1.300	[1.188, 1.437]	0.728	−1.617	2.781
1.240	[1.139, 1.360]	0.715	−2.287	3.824
1.260	[1.148, 1.395]	0.770	−1.689	3.708
1.212	[1.108, 1.337]	0.771	−1.742	3.405
1.301	[1.181, 1.449]	0.784	−1.744	3.322
1.286	[1.165, 1.435]	0.807	−2.134	4.235
1.328	[1.212, 1.469]	0.723	−1.571	3.414
1.235	[1.129, 1.361]	0.755	−1.764	3.464
1.221	[1.115, 1.349]	0.781	−2.417	3.125

[a] $1.21 \leqslant \sigma \leqslant 1.25$, $k = 400$, $n = 2500$, $m = 7.0$, $t = 20400$, observed coverage ≈94% (35 simulations).

the corresponding uniform distribution across this set. Here it is known that $\sigma \approx 1.8928$. It is possible to describe this situation with a three-dimensional mapping composed of three ternary shift maps; however, we used the method of iterated function systems (IFS) (see Barnsley and Demko[2]) to reproduce the attractor and distribution. A random vector (x, y, z) was generated in the unit cube and iteration performed by applying the "random map" $\mathbf{w}(x, y, z) = (w_1(x), w_2(y), w_3(z))$, where each of w_1, w_2, w_3 was chosen randomly and independently at each iteration from the set of two functions $\{u_1(x) = x/3,\ u_2(x) = 2/3 + x/3\}$. Since this IFS consists of contractive maps, it follows[24] that for all initial conditions this system eventually settles on to the intended attracting set; if the maps u_1 and u_2 are chosen with equal probability at each stage, then the distribution will be uniform. The result is a system which combines both deterministic and random components and possesses good mixing properties. The first 1000 iterates were discarded. Basepoints were separated by ten iterations (again a necessary separation) and the two samples by 50 iterations. Observations within each sample were taken successively. For smooth measures we have $[\theta(3, 7)]^{1/2} \approx 0.3107$. See Table X.

Table X. Example 4.8[a]

$\hat{\sigma}$	$[c_l, c_u]$	s	$R_{\min}$	$R_{\max}$
1.824	[1.672, 2.006]	0.498	−1.332	2.144
1.966	[1.797, 2.170]	0.479	−0.899	2.253
1.824	[1.673, 2.004]	0.494	−1.076	2.992
1.732	[1.594, 1.896]	0.499	−1.316	2.758
2.027	[1.842, 2.253]	0.495	−1.775	1.838
1.844	[1.690, 2.029]	0.495	−1.146	2.188
1.776	[1.642, 1.935]	0.461	−1.064	2.724
1.926	[1.763, 2.121]	0.479	−1.198	2.253
1.852	[1.707, 2.023]	0.458	−1.011	2.175
1.900	[1.731, 2.105]	0.513	−0.996	2.562
1.864	[1.702, 2.058]	0.508	−1.556	2.379
1.958	[1.803, 2.141]	0.437	−0.872	1.973
1.866	[1.711, 2.051]	0.485	−1.264	2.330
1.828	[1.686, 1.995]	0.460	−0.842	2.345
2.015	[1.831, 2.240]	0.498	−1.599	1.976
1.815	[1.662, 1.999]	0.507	−0.946	2.698
2.133	[1.934, 2.377]*	0.482	−0.996	2.068
1.887	[1.720, 2.090]	0.514	−1.748	2.316
1.954	[1.786, 2.157]	0.481	−1.199	2.359
1.810	[1.664, 1.983]	0.484	−0.640	2.704

[a] $\sigma = 1.8928$, $k = 400$, $n = 2500$, $m = 7.0$, $t = 20400$, observed coverage $\approx 95\%$ (25 simulations).

Example 4.9. A Fractal Cantor Distribution in $\mathbb{R}^3$. We used the IFS method of the preceding example to generate the same attractor $K \times K \times K$, but chose the maps u_1 and u_2 with probabilities 0.8 and 0.2, respectively, to produce a singular measure with $\sigma \approx 1.3665$. Note the increase in the observed values of s from those obtained in the previous example. See Table XI.

5. CONCLUDING REMARKS

In this paper we have tried to present an overview of the theory of dimension in dynamical systems, followed by a discussion of some statistical techniques for estimating the correlation and information dimensions. In Section 2 we developed a mathematical structure linking the various definitions of dimension and pointed out some open problems. In Section 3 we saw that the error in least squares estimates of the correlation dimension can be split into two components, a systematic error due to the inexact scaling of the attracting measure (shown to be a generic factor) and the statistical error due to sampling and estimation. Methods for coping

Table XI. Example 4.9[a]

$\hat{\sigma}$	$[c_l, c_u]$	s	$R_{\min}$	$R_{\max}$
1.370	[1.246, 1.520]	0.722	−1.670	3.380
1.219	[1.109, 1.354]*	0.817	−1.284	3.528
1.346	[1.209, 1.517]	0.840	−1.598	5.872
1.237	[1.126, 1.373]	0.798	−1.840	3.735
1.368	[1.241, 1.525]	0.751	−1.324	3.388
1.311	[1.183, 1.469]	0.820	−1.843	3.966
1.320	[1.181, 1.500]	0.891	−2.096	3.946
1.255	[1.142, 1.393]	0.790	−1.364	3.300
1.428	[1.271, 1.629]	0.864	−5.262	4.756
1.348	[1.212, 1.518]	0.833	−2.413	3.901
1.483	[1.339, 1.661]	0.726	−1.659	3.685
1.271	[1.153, 1.416]	0.804	−2.205	4.669
1.333	[1.208, 1.486]	0.774	−1.556	3.550
1.285	[1.161, 1.440]	0.836	−2.470	4.108
1.351	[1.211, 1.527]	0.856	−2.520	4.535
1.372	[1.233, 1.546]	0.823	−2.559	3.679
1.319	[1.194, 1.474]	0.794	−1.303	3.935
1.403	[1.243, 1.609]	0.916	−3.853	5.291
1.408	[1.267, 1.584]	0.790	−1.729	3.980
1.354	[1.231, 1.505]	0.740	−1.427	3.771

[a] $\sigma = 1.3665$, $k = 400$, $n = 2500$, $m = 7.0$, $t = 20400$, observed coverage ≈95% (25 simulations).

with both types of error were discussed, with particular success in the case of statistical error. In Section 4 we developed a nearest neighbor technique for constructing confidence intervals for the information dimension. This method appears to be an effective step toward establishing a statistical procedure with accurate error bounds. It is particularly appealing because it corrects for local density effect (thereby speeding up convergence in the case of highly nonuniform densities) and is applicable to a wide class of dynamical systems. It still remains to establish the appropriate rates of k_n and n in Theorem 4.1. Preliminary investigation suggests that the most important factor in determining the rate may be the true underlying dimension σ [the corresponding most important "error" appears to be the difference $|\mu_n - \mu|$ in (B.6) of Appendix B], although we can also expect the degree of "fractalness" of the measure to affect convergence. It is possible that the rate may be as bad as $k_n = o(n^{2/\sigma})$ for $\sigma \geqslant 2$; however, our numerical results indicate that almost satisfactory coverage can be obtained with reasonable sample sizes. This implies that the very large values of n required by a rate of $k_n = o(n^{2/\sigma})$ are only necessary "in the tail," i.e., if one insists on 95% coverage and is neither willing to accept a 93% or 92% coverage rate nor to widen the confidence interval slightly. It seems that satisfactory sample sizes and rates might be reasonably established numerically by exhaustive simulation of known systems. For example, the choices $k = 300$, $n = 500$, and $m = 5$ seem to perform well for most systems on the unit interval. However, it may not be worthwhile to determine sample sizes at this stage because it is clear that using higher-order nearest neighbors (such as second, third, and fourth nearest neighbors) will lead to a substantial improvement in our method. Higher-order nearest neighbors are more stable (Cutler[16] contains theoretical results on this) and will lead to shorter confidence intervals. We have had good success using higher-order nearest neighbors in the case of dynamical systems on the unit interval (generally reducing the width of intervals by 30–50%), but less success in higher dimensions, where the intervals often failed to cover, indicating inadequate sample sizes. Work on an improved procedure is in progress.

APPENDIX A. COMPLETION OF PROOF OF THEOREM 2.2

Let $\{r_n\}_n$ be any sequence of real numbers such that $r_n \downarrow 0$ and $\lim_{n \to \infty} (\log r_{n+1}/\log r_n) = 1$. Define $\sigma_n(x) = [\log \mathscr{m}(B(x, r_n))/\log r_n]$. For $r_{n+1} < r \leqslant r_n$ we obtain the inequalities

$$\left(\frac{\log r_n}{\log r}\right) \sigma_n(x) \leqslant \frac{\log \mathscr{m}(B(x, r))}{\log r} \leqslant \sigma_{n+1}(x) \left(\frac{\log r_{n+1}}{\log r}\right) \qquad \text{(A.1)}$$

which shows that the asymptotic behavior of $\log m(B(x, r))/\log r$ is determined by the behavior of the sequence of functions $\{\sigma_n\}_n$. We will need the following result.

Lemma A.1. Let C be a compact cube in $\mathbb{R}^N$ and m a probability measure on the Borel sets of C. Let $c(r)$ be any nonnegative function of r, and set $E(r) = \{x \in C \mid m(B(x, r)) \leqslant c(r)\}$. Then there exists a finite constant K_0 (depending only on N and the diameter of C) such that $m(E(r)) \leqslant K_0 c(r) r^{-N}$ whenever r does not exceed the diameter of C.

Proof. Since C is a compact cube in $\mathbb{R}^N$, there is a maximum number $K_0 r^{-N}$ of balls of radius r ($r \leqslant$ diameter of C) which can be centered at points of C in such a way that any two centers are at least the distance r apart. Since $E(r) \subseteq C$, we can find a covering $\mathscr{E}$ of $E(r)$ by balls of radius r with centers in $E(r)$ such that any two centers are separated by at least the distance r. It follows that $\mathscr{E}$ has at most $K_0 r^{-N}$ members, each of whose m-measure does not exceed $c(r)$. Consequently $m(E(r)) \leqslant K_0 c(r) r^{-N}$, as claimed. ∎

It will follow that

$$E\,(\limsup_{n\to\infty} \sigma_n(X)) \geqslant \limsup_{n\to\infty} E(\sigma_n(X))$$

[and also $E(\lim_{n\to\infty} \sigma_n(X)) = \lim_{n\to\infty} E(\sigma_n(X))$ in the case where $\sigma_P(x) = \sigma_H(x)$ m-a.s.] if we show that the functions $h_n(x) = \sup_{k\geqslant n} \sigma_k(x)$ are uniformly integrable. It is sufficient to check that $h_1(x) = \sup_{k\geqslant 1} \sigma_k(x)$ is integrable. Noting that

$$\begin{aligned} m(\{x \mid \sup_{k\geqslant 1} \sigma_k(x) > N+1+j\}) &= m\left(\bigcup_{k\geqslant 1} \{x \mid \sigma_k(x) > N+1+j\}\right) \\ &\leqslant \sum_{k=1}^{\infty} m(\{x \mid m(B(x, r_k)) < r_k^{N+1+j}\}) \\ &\leqslant K_0 \sum_{k=1}^{\infty} r_k^{N+1+j} r_k^{-N} = K_0 \sum_{k=1}^{\infty} r_k^{1+j} \end{aligned}$$

(using Lemma A.1), we obtain

$$E(h_1) \leqslant (N+1) + \sum_{j=0}^{\infty} (N+j+2)\, E(I_{[N+1+j<h_1\leqslant N+j+2]})$$

$$\begin{aligned}
&\leqslant (N+1) + \sum_{j=0}^{\infty} (N+j+2)\, \mathscr{m}(\{x \mid \sup_{k \geqslant 1} \sigma_k(x) > N+1+j\}) \\
&\leqslant (N+1) + \sum_{j=0}^{\infty} (N+j+2) K_0 \sum_{k=1}^{\infty} r_k^{1+j} \\
&\leqslant (N+1) + K_0 \sum_{k=1}^{\infty} \sum_{j=0}^{\infty} (N+j+2) r_k^{1+j} \\
&= (N+1) + K_0 \sum_{k=1}^{\infty} r_k \left(\frac{N+1}{1-r_k} + \frac{1}{(1-r_k)^2} \right) \\
&\leqslant (N+1) + K_0^* \sum_{k=1}^{\infty} r_k \qquad \text{for some constant } K_0^* \\
&= (N+1) + K_0^* \qquad \text{choosing } r_k = 2^{-k}
\end{aligned}$$

Hence h_1 is integrable. This completes the proof of Theorem 2.2. ▮

APPENDIX B. A CENTRAL LIMIT THEOREM FOR AN ARRAY OF EXCHANGEABLE VARIABLES

As noted earlier, the following theorems are generalizations of problems #3, 4, p. 317 in Chow and Teicher.[12] A special case of these results can be found in Blum *et al.*[9]

Theorem B.1. Let $Y_{n,j}$, $n = 1, 2,...$ and $j = 1, 2,..., k_n$, denote an array of random variables such that the variables within each row can be embedded in an infinite exchangeable sequence. Suppose there exists a common set of random variables Q such that $Y_{n,1}, Y_{n,2},..., Y_{n,k_n}$ are conditionally i.i.d. (independent and identically-distributed) when conditioned on Q. Let $\mu_n(Q) = E(Y_{n,j} \mid Q)$ and $\theta_n(Q) = \mathrm{Var}(Y_{n,j} \mid Q) = E((Y_{n,j} - \mu_n(Q))^2 \mid Q)$. Let P denote the joint distribution of all the $Y_{n,j}$ and the variables in Q, and let P^* denote the distribution of the variables in Q alone. Suppose $k_n \to \infty$ as $n \to \infty$ and the following three conditions are satisfied:

There exists a value μ such that $k_n^{1/2}(\mu_n(Q) - \mu) \xrightarrow{P^*} 0$

$$\text{as} \quad n \to \infty \tag{B.1}$$

$$\text{There exists } \theta > 0 \text{ such that } \lim_{n\to\infty} P^*(\theta_n(Q) > \theta) = 1 \tag{B.2}$$

$$\Gamma_{n,k_n,Q} = k_n^{-1/2}\theta_n(Q)^{-3/2} E(|Y_{n,j} - \mu_n(Q)|^3 \mid Q) \xrightarrow{P^*} 0 \quad \text{as} \quad n \to \infty \tag{B.3}$$

Then

$$Z_{n,k_n} = \frac{\sum_{j=1}^{k_n} Y_{n,j} - k_n\mu}{[k_n\theta_n(Q)]^{1/2}}$$

converges in distribution to a standard Gaussian distribution with mean 0 and variance 1 as $n \to \infty$.

Proof. We first note that the exchangeability assumption guarantees for each n the existence of a set Q_n such that the variables in the nth row become i.i.d. when conditioned on Q_n, and in fact this theorem can be proven in this more general case. However, for our purposes it will be sufficient and notationally simpler to assume the existence of a common set Q which works for all n. Now write

$$Z_{n,k_n} = \frac{\sum_{j=1}^{k_n} Y_{n,j} - k_n\mu_n(Q)}{[k_n\theta_n(Q)]^{1/2}} + \frac{k_n(\mu_n(Q) - \mu)}{[k_n\theta_n(Q)]^{1/2}} \tag{B.4}$$

Note that the second term on the rhs of Eq. (B.4) tends to 0 in probability as $n \to \infty$ if conditions (B.1) and (B.2) hold. Hence we need only show that the first term on the rhs, which we will denote by $Z_{n,k_n,Q}$, converges in distribution to a Gaussian. Let $\Phi(y) = \int_{-\infty}^{y} e^{-t^2/2}(2\pi)^{-1/2}\,dt$ and define $D_n(q, y) = |P(Z_{n,k_n,Q} \leqslant y \mid Q = q) - \Phi(y)|$. Then we have

$$|P(Z_{n,k_n,Q} \leqslant y) - \Phi(y)| \leqslant \int D_n(q, y)\, P^*(dq) \tag{B.5}$$

Now by the Berry–Esseen inequality for independent variables, there exists a universal constant C such that $D_n(q, y) \leqslant C\Gamma_{n,k_n,q}$ for all y and P^*-almost all q. Since $D_n(q, y)$ is bounded by 2, it follows from the bounded convergence theorem that the rhs of (B.5) tends to 0 as $n \to \infty$ if condition (B.3) holds. Hence the theorem is proved. ▮

We now need to replace conditions (B.1)–(B.3) by conditions which are more easily verified in practice. We obtain the following:

Theorem B.2. Let $Y_{n,j}$, $n = 1, 2,\ldots$ and $j = 1, 2,\ldots, k_n$, denote an array of exchangeable random variables with common set Q as in

Theorem B.1. Let $\mu_n(Q)$, $\theta_n(Q)$, and $\Gamma_{n,k_n,Q}$ be defined as in Theorem B.1. Set $\mu_n = E(\mu_n(Q)) = E(Y_{n,j})$. Suppose the following conditions are satisfied:

$$\begin{aligned} &\operatorname{Cov}(Y_{n,i}, Y_{n,j}) = o(k_n^{-1}),\\ &\quad\text{and there exists a value } \mu \text{ such that}\\ &\quad |\mu_n - \mu| = o(k_n^{-1/2}) \end{aligned} \tag{B.6}$$

$$\begin{aligned} &\text{There exists } \theta > 0\\ &\quad\text{and a sequence } \{a_n\}_n \text{ of positive constants}\\ &\quad\text{with } a_n \to a,\ 0 < a \leqslant \infty, \text{ such that}\\ &\quad\operatorname{Cov}((Y_{n,i}/a_n)^2, (Y_{n,j}/a_n)^2) = o(1)\\ &\quad\text{and } \operatorname{Var}(Y_{n,j}/a_n) \to \theta \end{aligned} \tag{B.7}$$

$$\begin{aligned} &E(|Y_{n,j}/a_n - \mu_n/a_n|^3) = o(k_n^{1/2})\\ &\quad\text{for the sequence } \{a_n\}_n \text{ defined}\\ &\quad\text{in (B.7)} \end{aligned} \tag{B.8}$$

Then conditions (B.1)-(B.3) hold and the conclusions of Theorem B.1 follow.

Proof. Note that we may partition $E(k_n(\mu_n(Q) - \mu)^2)$ as follows:

$$\begin{aligned} E(k_n(\mu_n(Q) - \mu)^2) &= k_n(\mu_n - \mu)^2 + E(k_n(\mu_n(Q) - \mu_n)^2)\\ &= k_n(\mu_n - \mu)^2 + k_n \operatorname{Var}(\mu_n(Q)) \end{aligned} \tag{B.9}$$

From (B.6) it follows immediately that the first term on the rhs of (B.9) tends to 0. To see that the second term also tends to 0, consider that

$$\begin{aligned} \operatorname{Cov}(Y_{n,i}, Y_{n,j}) &= E(\operatorname{Cov}(Y_{n,i}, Y_{n,j} \mid Q))\\ &\quad + \operatorname{Cov}(E(Y_{n,i} \mid Q), E(Y_{n,j} \mid Q)) \end{aligned} \tag{B.10}$$

Now the first term on the rhs of (B.10) vanishes because of the conditional independence of $Y_{n,i}$ and $Y_{n,j}$ given Q, and we see that $\operatorname{Cov}(E(Y_{n,i} \mid Q), E(Y_{n,j} \mid Q)) = \operatorname{Var}(\mu_n(Q))$. Hence from (B.6) we conclude $E(k_n(\mu_n(Q) - \mu)^2) \to 0$, which certainly implies (B.1). Now let $\{a_n\}_n$ be the sequence in (B.7), and define $\tilde{Y}_{n,j} = Y_{n,j}/a_n$, $\tilde{\theta}_n(Q) = \theta_n(Q)/a_n^2$, and $\tilde{\mu}_n(Q) = \mu_n(Q)/a_n$. Let θ be as in (B.7) and note that if $\tilde{\theta}_n(Q) \xrightarrow{P^*} \theta$, then (B.2) must follow (with some $\tilde{\theta} < \theta$ in the role of θ in that equation). Hence, to obtain (B.2) it suffices to show that $\tilde{\theta}_n(Q) \xrightarrow{P^*} \theta$. Now write

$$\tilde{\theta}_n(Q) = E(\tilde{Y}_{n,i}^2 \mid Q) - \tilde{\mu}_n(Q)^2 \tag{B.11}$$

From the earlier work in this proof we know that $\tilde{\mu}_n(Q)^2 \xrightarrow{P^*} \mu^2/a^2$, where we define $\mu^2/a^2 = 0$ if $a = \infty$. Now consider the first term on the rhs of (B.11). We see that

$$E(E(\tilde{Y}^2_{n,j} \mid Q)) = E(\tilde{Y}^2_{n,j}) = \mu^2_n/a^2_n + \operatorname{Var}(\tilde{Y}_{n,j}) \to \mu^2/a^2 + \theta$$

from (B.6), (B.7). Expressing $\operatorname{Cov}(\tilde{Y}^2_{n,i}, \tilde{Y}^2_{n,j})$ by conditioning on Q, as in (B.10), yields

$$\operatorname{Var}(E(\tilde{Y}^2_{n,j} \mid Q)) = \operatorname{Cov}(\tilde{Y}^2_{n,i}, \tilde{Y}^2_{n,j}) \to 0$$

by (B.7). Thus, we conclude $E(\tilde{Y}^2_{n,j} \mid Q) \xrightarrow{P^*} \mu^2/a^2 + \theta$, giving the desired result $\tilde{\theta}_n(Q) \xrightarrow{P^*} \theta$. It remains only to verify that (B.3) holds. Note we have the equality $\Gamma_{n,k_n,Q} = \tilde{\Gamma}_{n,k_n,Q}$ where we define

$$\tilde{\Gamma}_{n,k_n,Q} = k_n^{\ 1/2} \tilde{\theta}_n(Q)^{\ 3/2}\, E(|\tilde{Y}_{n,j} - \tilde{\mu}_n(Q)|^3 \mid Q) \tag{B.12}$$

Hence it is sufficient to verify (B.3) for $\tilde{\Gamma}_{n,k_n,Q}$. Now, since $\tilde{\theta}_n(Q) \xrightarrow{P^*} \theta$, we need only consider the variable in the numerator of (B.12). Let $\tilde{\mu}_n = \mu_n/a_n$, and define $h_n(Q) = |\tilde{\mu}_n - \tilde{\mu}_n(Q)|$ and $T_n = |\tilde{Y}_{n,j} - \tilde{\mu}_n|$. We have the following inequality:

$$\begin{aligned}\tilde{\theta}_n(Q)^{3/2}\, \tilde{\Gamma}_{n,k_n,Q} \leqslant k_n^{\ 1/2} E(T_n^3 \mid Q) + 3k_n^{\ 1/2} h_n(Q)\, E(T_n^2 \mid Q) \\ + 3k_n^{-1/2} h_n(Q)^2\, E(T_n \mid Q) + k_n^{\ 1/2} h_n(Q)^3\end{aligned} \tag{B.13}$$

Now consider the first term on the rhs of (B.13). From (B.8) we see that

$$k_n^{-1/2} E(E(T_n^3 \mid Q)) = k_n^{-1/2} E(T_n^3) \to 0$$

and hence we must have $k_n^{-1/2} E(T_n^3 \mid Q) \xrightarrow{P^*} 0$. Now $h_n(Q) \xrightarrow{P^*} 0$ and

$$\max(E(T_n \mid Q), E(T_n^2 \mid Q)) \leqslant 1 + E(T_n^3 \mid Q)$$

so it follows that the remaining terms on the rhs of (B.13) also tend to 0 in probability. This shows that (B.3) holds and completes the proof of the theorem. ▌

APPENDIX C. APPLYING THE CENTRAL LIMIT THEOREM TO NEAREST NEIGHBORS

In this Appendix we justify Theorem 4.1 by arguing that the conditions (B.6)–(B.8) of Theorem B.2 will be met by a variety of distributions $\mathscr{m}$. Set $Y_{n,j} = R_n(X_j)$, define $Q = \bigcup_{n=1}^{\infty} S_n$, and note that $\theta(S_n) = \theta_n(Q)$ [where $\theta(S_n)$ is defined in Theorem 4.1 and $\theta_n(Q)$ is defined in

Theorem B.1]. We first show, under a very weak condition on $\mathscr{m}$ (which, surprisingly, involves the correlation dimension), that all moments of $R_n(X_j)$ exist [and hence the conditional moments $\mu_n(Q)$ and $\theta_n(Q)$ exist and are finite P^*-a.s.]. Since

$$R_n(X_j) = \frac{1}{\log m} (\log d_{1,n}(X_j) - \log d_{2,mn}(X_j))$$

it is sufficient to consider the moments of $\log d_{1,n}(X_j)$. Noting that

$$\log d_{1,n}(X_j) = \min_{1 \leqslant u \leqslant n} \log \| W_{1,u} - X_j \|$$

$$E(| \min_{1 \leqslant u \leqslant n} \log \| W_{1,u} - X_j \| |) \leqslant nE(|\log \| W_{1,u} - X_j \| |)$$

we see that the problem reduces to considering the moments of $\log \| W - X \|$, where W and X are two independent observations from $\mathscr{m}$.

Lemma C.1. Let $\mathscr{m}$ be a probability distribution on a compact subset of $\mathbb{R}^N$ such that the lower correlation dimension $\nu \;\; = m \;\; (1) > 0$. Then all moments of $\log \| W - X \|$ exist.

Proof. Let $C(r)$ denote the correlation integral defined in (2.17). If $\nu \;\; > 0$, it follows that $\lim_{r \to 0} (\log r)^n r^{\;\;\delta} C(r) = 0$ for every $0 < \delta < \nu$ and every positive integer n. Now consider that we may write

$$E(|\log \| W - X \| |^n) = E(|\log \| W - X \| |^n I_{[\| W - X \| \geqslant 1]})$$
$$+ E(|\log \| W - X \| |^n I_{[\| W - X \| < 1]}) \qquad \text{(C.1)}$$

The first term on the rhs of (C.1) is bounded because of the compactness of the support of $\mathscr{m}$. Now consider the second term, and note that

$$E(|\log \| W - X \| |^n I_{[\| W - X \| < 1]}) = \left| \int_0^1 (\log r)^n C(dr) \right| \qquad \text{(C.2)}$$

Consider first $n = 1$. Integrating by parts yields

$$\int_0^1 (\log r)\, C(dr) = - \int_0^1 C(r)/r \, dr = - \int_0^1 r^{\delta \;\; 1} C(r)/r^\delta \, dr \qquad \text{(C.3)}$$

for any $0 < \delta < \nu^-$. As $C(r)/r^\delta$ stays bounded as $r \to 0$ and $r^{\delta - 1}$ is integrable, it follows that the first moment of $\log \| W - X \|$ is finite. Higher moments $(n \geqslant 2)$ follow by a similar argument. ∎

Thus, all quantities (conditional and unconditional moments) defined in Appendix B do exist in our application. It remains to indicate how

(B.6)–(B.8) are met. It is not possible to provide detailed proofs here, but we indicate the methods of proof and refer to results established in earlier papers. First note that we can write

$$R_n(X_j) = \frac{L_{2,mn}(X_j) - L_{1,n}(X_j)}{\sigma \log m} + \frac{1}{\sigma} \tag{C.4}$$

where

$$L_{1,n}(X_j) = \sigma \log n \left(\frac{\log d_{1,n}(X_j)}{\log 1/n} - \frac{1}{\sigma} \right)$$

and $L_{2,mn}(X_j)$ is defined analogously for the second sample. Cutler and Dawson[19,20] considered the asymptotic behavior of $L_n(X)$ in various situations. We summarize the relevant points and conclusions below:

1. *Smooth Measures.* If m is a smooth measure in $\mathbb{R}^N$, then for m-almost all x, $L_n(x)$ converges in distribution to $\mathrm{EV}(\log(K_\sigma g(x)), 1)$ where $g(x)$ is the density of m at x, K_σ is a normalizing constant depending only on σ, and $\mathrm{EV}(a, b)$ denotes an extreme value distribution with location parameter a and scale parameter b. The distribution function of $\mathrm{EV}(a, b)$ is given by $F(y) = \exp\{-\exp[-(y-a)/b]\}$ for $-\infty < y < \infty$, and the mean and variance of $\mathrm{EV}(a, b)$ are $\gamma + a$ and $b^2\pi^2/6$ respectively (γ = Euler's constant). Convergence of all moments is also proven for a very wide class of smooth measures (in fact, by appealing to results of Pickands,[47] it would appear we can extend this class to essentially all smooth measures). It follows that for a random basepoint X, $L_n(X)$ converges in distribution to a compounded extreme value distribution with location parameter $E(\log K_\sigma g(X))$ and scale parameter 1. Under mild restrictions we can expect the moments of $L_n(X)$ to also converge. (One necessary restriction, obviously, is that $E(\log g(X))$ be finite.) As a consequence the asymptotic mean of $L_n(X)$ is $\gamma + E(\log K_\sigma g(X))$ and the asymptotic variance is $\pi^2/6$. The limit theory of Theorem B.2 is applicable, taking $a_n = 1$ and $R_n(X_j) = Y_{n,j} = \tilde{Y}_{n,j}$. We have $\mu_n = E(R_n(X_j)) \to 1/\sigma$ and $\mathrm{Var}(R_n(X_j)) \to \pi^2/3(\log m)^2\,\sigma^2$ [both as a consequence of (C.4) and the moment convergence of $L_{1,n}(X_j)$ and $L_{2,mn}(X_j)$]. Condition (B.8) follows (for any increasing rate k_n) because of third moment convergence. It remains only to argue that the covariances in (B.6) and (B.7) vanish as $n \to \infty$ (the actual rate k_n to be applied in the problem as a whole is determined by considering the various rates associated with the relevant quantities and selecting a rate which works for all). But it is not difficult to show that $R_n(X_i)$ and $R_n(X_j)$ are asymptotically independent, and the convergence of moments therefore implies that the covariances vanish in the limit. [Bickel and Breiman[5] established a vanishing rate of $o(1/n)$

for covariances of certain functions of nearest neighbors in the case of absolutely continuous measures.] Thus, (B.6)–(B.8) will hold for some rate k_n. As noted in the concluding remarks, we expect the rate at which $|E(R_n(X)) - 1/\sigma| \to 0$ to be the most important factor in determining rates and sample sizes.

2. *Fractal Measures.* While the distribution of $L_n(x)$ fails to converge in the fractal case, the damped quantity $L_n(X)/\sigma(\log n)^{1/2}$ (where X is a randomly selected basepoint) asymptotically follows a Gaussian with mean 0 and some variance $\theta > 0$. Under mild restrictions, moments also converge. This gives $\mathrm{Var}(R_n(X)) \approx 2/(\log m)^2\, \theta(\log n)$ and it can be seen from the form of θ (given in Cutler and Dawson[20]) that θ is larger when m is "more singular" or "more fractal." The limit theory of Theorem B.2 is applicable here with $a_n = (\log n)^{1/2}$, $R_n(X_j) = Y_{n,j}$, and, consequently, $\tilde{Y}_{n,j} = R_n(X_j)/(\log n)^{1/2}$. We conclude that (B.7) and (B.8) hold and $\mathrm{Cov}(R_n(X_i), R_n(X_j)) \to 0$, so it remains only to argue that $E(R_n(X)) \to 1/\sigma$. From (C.4) it is equivalent to show that $E(L_{2,mn}(X) - L_{1,n}(X)) \to 0$. Setting $g(n) = E(L_n(X))$, we note that $g(n) = o((\log n)^{1/2})$. Hence, provided that $g(n)$ does not have a nonvanishing oscillatory term (this seems a reasonable assumption), then $|g(n) - g(mn)| \to 0$ as required. [We may actually have $g(n)$ converging as $n \to \infty$, but rigorous results do not yet exist on this.] Numerically, we do observe $\bar{R}_n$ centering around $1/\sigma$ over repeated simulations.

3. *Semifractal Measures.* Here, for fixed basepoints, $L_n(x)$ oscillates between two extreme value distributions [whose centers are determined by the density $g(x)$ and two constants reflecting the geometry of the supporting Cantor set] and moments stay bounded as $n \to \infty$. We expect that we may proceed as in the smooth case, taking $a_n = 1$, under the assumption $E(L_{2,mn}(X) - L_{1,n}(X)) \to 0$.

ACKNOWLEDGMENTS

I am grateful to the referees for several useful suggestions. This research was supported by the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

1. R. Badii and A. Politi, Statistical description of chaotic attractors: The dimension function, *J. Stat. Phys.* **40**:725–750 (1985).
2. M. F. Barnsley and S. Demko, Iterated function systems and the global construction of fractals, *Proc. R. Soc. Lond. A* **399**:243–275 (1985).

3. C. Beck, Upper and lower bounds on the Renyi dimensions and the uniformity of multifractals, *Physica D* **41**:67–78 (1990).
4. R. Benzi, G. Paladin, G. Parisi, and A. Vulpiani, On the multifractal nature of fully developed turbulence and chaotic systems, *J. Phys. A: Math. Gen.* **17**:3521–3531 (1984).
5. P. J. Bickel and L. Breiman, Sums of functions of nearest neighbor distances, moment bounds, limit theorems, and a goodness-of-fit test, *Ann. Prob.* **11**:185–214 (1983).
6. P. Billingsley, Hausdorff dimension in probability theory, *Illinois J. Math.* **4**:187–209 (1960).
7. P. Billingsley, Hausdorff dimension in probability theory II, *Illinois J. Math.* **5**:291–298 (1961).
8. P. Billingsley, *Ergodic Theory and Information* (Krieger, New York, 1978).
9. J. R. Blum, H. Chernoff, M. Rosenblatt, and H. Teicher, Central limit theorems for interchangeable processes, *Can. J. Math.* **10**:222–229 (1958).
10. R. Bowen and D. Ruelle, The ergodic theory of Axiom A flows, *Invent. Math.* **29**:181 202 (1975).
11. R. Cawley and R. D. Mauldin, Multifractal decompositions of Moran fractals, Preprint (1990).
12. Y. S. Chow and H. Teicher, *Probability Theory: Independence, Interchangeability, Martingales*, 2nd ed. (Springer-Verlag, New York, 1988).
13. P. Collet, J. L. Lebowitz, and A. Porzio, The dimension spectrum of some dynamical systems, *J. Stat. Phys.* **47**:609–644 (1987).
14. C. D. Cutler, The Hausdorff dimension distribution of finite measures in Euclidean space, *Can. J. Math.* **38**:1459–1484 (1986).
15. C. D. Cutler, Connecting ergodicity and dimension in dynamical systems, *Ergodic Theory Dynam. Syst.* **10**:451–462 (1990).
16. C. D. Cutler, kth nearest neighbors and the generalized logistic distribution, in *The Logistic Distribution*, N. Balakrishnan, ed. (Marcel Dekker, New York, 1991).
17. C. D. Cutler, A dynamical system with integer information dimension and fractal correlation exponent, *Commun. Math. Phys.* **129**:621–629 (1990).
18. C. D. Cutler, Measure disintegrations with respect to σ-stable monotone indices and the pointwise representation of packing dimension, Preprint (1990).
19. C. D. Cutler and D. A. Dawson, Estimation of dimension for spatially-distributed data and related limit theorems, *J. Multivariate Anal.* **28**:115–148 (1989).
20. C. D. Cutler and D. A. Dawson, Nearest neighbor analysis of a family of fractal distributions, *Ann. Prob.* **18**:256–271 (1990).
21. M. Denker and G. Keller, Rigorous statistical procedures for data from dynamical systems, *J. Stat. Phys.* **44**:67–93 (1986).
22. B. Dubuc, J. F. Quiniou, C. Roques-Carmes, C. Tricot, and S. W. Zucker, Evaluating the fractal dimension of profiles, *Phys. Rev. A* **39**:1500–1512 (1989).
23. J.-P. Eckmann and D. Ruelle, Ergodic theory of chaos and strange attractors, *Rev. Mod. Phys.* **57**:617–656 (1985).
24. J. H. Elton, An ergodic theorem for iterated maps, *Ergodic Theory Dynam. Syst.* **7**:481–488 (1987).
25. C. Essex, Correlation dimension and data sample size, Preprint (1987).
26. C. Essex, T. Lookman, and M. A. H. Nerenberg, The climate attractor over short time scales, *Nature* **326**:64–66 (1987).
27. K. J. Falconer, *The Geometry of Fractal Sets* (Cambridge University Press, Cambridge, 1985).
28. J. D. Farmer, E. Ott, and J. A. Yorke, The dimension of chaotic attractors, *Physica D* **7**:153–180 (1983).

29. P. Grassberger, Are there really climate attractors?, *Nature* **322**:609–612 (1986).
30. P. Grassberger and I. Procaccia, Characterization of strange attractors, *Phys. Rev. Lett.* **50**:346–349 (1983).
31. P. Grassberger and I. Procaccia, Measuring the strangeness of strange attractors, *Physica D* **9**:189–208 (1983).
32. P. Grassberger, R. Badii, and A. Politi, Scaling laws for invariant measures on hyperbolic and nonhyperbolic attractors, *J. Stat. Phys.* **51**:135–178 (1988).
33. H. S. Greenside, A. Wolf, J. Swift, and T. Pignataro, Impracticality of a box-counting algorithm for calculating the dimensionality of strange attractors, *Phys. Rev. A* **25**:3453–3456 (1982).
34. J. Guckenheimer, Dimension estimates for attractors, *Contemp. Math.* **28**:357–367 (1984).
35. T. C. Halsey, M. H. Jensen, L. P. Kadanoff, I. Procaccia, and B. I. Shraiman, Fractal measures and their singularities: The characterization of strange sets, *Phys. Rev. A* **33**:1141–1151 (1986).
36. M. Hénon, A two-dimensional mapping with a strange attractor, *Commun. Math. Phys.* **50**:69–77 (1976).
37. H. G. E. Hentschel and I. Procaccia, The infinite number of generalized dimensions of fractals and strange attractors, *Physica D* **8**:435–444 (1983).
38. J. Holzfuss and G. Mayer-Kress, An approach to error estimation in the application of dimension algorithms, in *Dimensions and Entropies in Chaotic Systems: Quantification of Complex Behavior*, G. Mayer-Kress, ed. (Springer-Verlag, New York, 1986).
39. F. Ledrappier and M. Misiurewicz, Dimension of invariant measures for maps with exponent zero, *Ergodic Theory Dynam. Syst.* **5**:595–610 (1985).
40. G. Mayer-Kress, ed., *Dimensions and Entropies in Chaotic Systems: Quantification of Complex Behavior* (Springer-Verlag, New York, 1986).
41. R. H. Myers, *Classical and Modern Regression with Applications* (Duxbury Press, Boston, 1986).
42. C. Nicolis and G. Nicolis, Is there a climatic attractor?, *Nature* **311**:529–532 (1984).
43. E. Ott, W. D. Withers, and J. A. Yorke, Is the dimension of chaotic attractors invariant under coordinate changes?, *J. Stat. Phys.* **36**:687-697 (1984).
44. N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw, Geometry from a time series, *Phys. Rev. Lett.* **45**:712–716 (1980).
45. G. Paladin and S. Vaienti, Hausdorff dimensions in two-dimensional maps and thermodynamic formalism, *J. Stat. Phys.* **57**:289–299 (1989).
46. G. Paladin and A. Vulpiani, Anomalous scaling laws in multifractal objects, *Phys. Rep.* **156**:147-225 (1987).
47. J. Pickands III, Moment convergence of sample extremes, *Ann. Math. Stat.* **39**:881–889 (1968).
48. J. B. Ramsey and H.-J. Yuan, The statistical properties of dimension calculations using small data sets, *Nonlinearity* **3**:155–176 (1990).
49. D. A. Rand, The singularity spectrum $f(\alpha)$ for cookie-cutters, *Ergodic Theory Dynam. Syst.* **9**:527–541 (1989).
50. C. A. Rogers, *Hausdorff Measures* (Cambridge University Press, Cambridge, 1970).
51. W. Rudin, *Real and Complex Analysis*, 2nd ed. (McGraw-Hill, New York, 1974).
52. X. Saint Raymond and C. Tricot, Packing regularity of sets in n-space, *Math. Proc. Camb. Phil. Soc.* **103**:133–145 (1988).
53. F. Takens, Detecting strange attractors in turbulence, in *Lecture Notes in Mathematics*, 898 (Springer-Verlag, Berlin, 1981), pp. 366–381.
54. F. Takens, On the numerical determination of the dimension of an attractor, in *Lecture Notes in Mathematics*, 1125 (Springer-Verlag, Berlin, 1985), pp. 99–106.

55. S. J. Taylor, The measure theory of random fractals, *Math. Proc. Camb. Phil. Soc.* **100**:383–406 (1986).
56. C. C. Taylor and S. J. Taylor, Estimating the dimension of a fractal, *J. R. Stat. Soc.*, to appear.
57. S. J. Taylor and C. Tricot, Packing measure, and its evaluation for a Brownian path, *Trans. Am. Math. Soc.* **288**:679–699 (1985).
58. S. J. Taylor and C. Tricot, The packing measure of rectifiable subsets of the plane, *Math. Proc. Camb. Phil. Soc.* **99**:285–296 (1986).
59. Y. Termonia and Z. Alexandrowicz, Fractional dimension of strange attractors from radius versus size of arbitrary clusters, *Phys. Rev. Lett.* **51**:1265–1268 (1983).
60. C. Tricot, Rarefaction indices, *Mathematika* **27**:46–57 (1980).
61. C. Tricot, Two definitions of fractional dimension, *Math. Proc. Camb. Phil. Soc.* **91**:57–74 (1982).
62. C. Tricot, J. F. Quiniou, D. Wehbi, C. Roques-Carmes, and B. Dubuc, Evaluation de la dimension fractale d'un graphe, Preprint (1987).
63. L.-S. Young, Dimension, entropy, and Lyapunov exponents, *Ergodic Theory Dynam. Syst.* **2**:109–124 (1982).