Comment: Relation Between Statistics and Chaos

Richard L. Smith

The involvement of statisticians in the field of chaos is relatively recent, but rapidly growing. Howell Tong's book (Tong, 1990) did much to make statisticians aware of the field. The Royal Statistical Society has hosted discussion papers by Bartlett (1990), and papers from a recent one-day meeting will appear in a special issue of the *Journal of the Royal Statistical Society, Series B* in 1992. I am delighted to see that *Statistical Science* is also taking a lead in developing this fertile source of statistical problems.

Both of the articles under discussion, Berliner (1992) and Chatterjee and Yilmaz (1992), are essentially expository, outlining the theory of chaos in a manner that is oriented toward statistical application. Berliner's article in particular shows how notions of ergodic theory, which tends to be regarded as being "at the hard end" of deterministic dynamical systems theory, have simple probabilistic interpretations that make the theory appealing to statisticians, even though it is essentially describing deterministic systems.

In developing some more specific themes on which I can comment in some detail, I would like to concentrate particularly on the contribution that statisticians can make to the interpretation of data from dynamical systems.

There is an extensive literature on the mathematical properties of systems, such as the logistic map or Lorenz's system of differential equations, and there are also physical systems such as Taylor–Couette flow where the underlying dynamics of the system is sufficiently well understood for a direct association to be made between mathematical theory and experimental observation. But in areas such as ecology or economics, it is impossible to know the detailed mathematical equations governing the system, and the whole of the evidence for "chaos," if indeed there is any evidence at all, comes from the interpretation of experimental data.

Richard L. Smith is Professor of Statistics, University of North Carolina at Chapel Hill, CB 3260, Phillips Hall, Chapel Hill, North Carolina 27599-3260.

Ruelle (1990) gave a particularly witty demonstration of how easy it is to misinterpret such data.

I will focus on just one of the numerous techniques proposed, namely the estimation of correlation dimension. Suppose we have a univariate time series $\{X_n\}$, and form d-dimensional embedded vectors $\mathbf{Y}_n = (X_{n-d+1}, X_{n-d+2}, \ldots, X_n)$ or more generally $\mathbf{Y}_n = (X_{n-(d-1)\tau}, X_{n-(d-2)\tau}, \ldots, X_n)$, where τ is a *lag parameter*. Define

$$C_N(r) = \frac{\sum_{i=2}^{N} \sum_{j=1}^{i} I(\|\mathbf{Y}_i - \mathbf{Y}_j\| < r)}{N(N-1)/2}$$

where N is the number of observations, I denotes the indicator function and $\|\cdot\|$ is a norm. The limit

$$C(r) = \lim_{N\to\infty} C_N(r)$$

is called the *correlation integral*. The correlation dimension, when it is defined, is given by

$$\nu = \lim_{r\to 0} \frac{\log C(r)}{\log r}. \tag{1}$$

In the context of fractals, these formulas give a relatively straightforward way of determining a dimension of a fractal. In the context of chaotic time series, if it is possible to estimate a correlation dimension, which for large enough embedding dimension d is independent of d within the limits of statistical error, then this is often taken as an indicator of deterministic chaos as opposed to randomness.

Most current algorithms for calculating ν from experimental or observational data essentially consist of regressing $\log C_N(r)$ on $\log r$ over a suitable range of r. An alternative technique is the following. First, we strengthen (1) to

$$C(r) \sim ar^{\nu} \quad \text{as } r \downarrow 0. \tag{2}$$

This gives an asymptotic power-law tail for the distribution of distance between two arbitrarily chosen points of the attractor. In practice, we may choose to simplify this even further to

$$C(r) = ar^{\nu} \quad \text{for } r < \varepsilon \tag{3}$$

for some threshold ε, which will be considered further below. A second assumption is that the