A review of symbolic analysis of experimental data

C.S. Daw and C.E.A. Finney

Oak Ridge National Laboratory, Knoxville, Tennessee 37932-6472

E.R. Tracy

College of William and Mary, Williamsburg, Virginia 23187-8795
(Dated: 2002-07-22.)

This review covers the group of data-analysis techniques collectively referred to as symbolization or symbolic time-series analysis. Symbolization involves transformation of raw time-series measurements (*i.e.*, experimental signals) into a series of discretized symbols that are processed to extract information about the generating process. In many cases, the degree of discretization can be quite severe, even to the point of converting the original data to single-bit values. Current approaches for constructing symbols and detecting the information they contain are summarized. Novel approaches for characterizing and recognizing temporal patterns can be important for many types of experimental systems, but this is especially true for processes that are nonlinear and possibly chaotic. Recent experience indicates that symbolization can increase the efficiency of finding and quantifying information from such systems, reduce sensitivity to measurement noise, and discriminate both specific and general classes of proposed models. Examples of the successful application of symbolization to experimental data are included. Key theoretical issues and limitations of the method are also discussed.

I. INTRODUCTION

Experiments involving dynamic measurements typically require careful definition of the physical quantities to be measured and the instrumental means by which the measurements will be made. One is often interested in testing hypotheses or making inferences on the basis of temporal patterns in time-series data. When the observed dynamics are relatively simple, such as sinusoidal periodicities, traditional analytical tools such as Fourier transforms are easily used to characterize the patterns. More complex dynamics, such as bifurcations and chaotic oscillations, can require more sophisticated approaches. In the latter case especially, the method of data analysis should be selected with careful consideration for the experimental setup and the underlying physics (if they are known). Details such as the dynamic instrument response, the digital sampling rate, and the signal-to-noise ratio can significantly affect the reliability of the results.

Our objective in this review is to summarize recent developments in the application of a data-analysis technique referred to as symbolization or symbolic time-series analysis. A central step in the technique is discretizing the raw time-series measurements into a corresponding sequence of symbols. The symbol sequence is then treated as a transform for the original data that retains much of the important temporal information. An important practical advantage of working with symbols is that the efficiency of numerical computations is greatly increased over what it would be for the original data. In some cases, efficiency may be mainly of value for reducing the need for computational resources or enhancing understanding, but it can also imply speed as well. The latter may be important for real-time monitoring and control applications.[1] Also, analysis of symbolic data is often less sensitive to measurement noise. In some cases,

symbolization can be accomplished directly in the instrument by appropriate design of the sensing elements. Such low-resolution (even "disposable") sensors combined with appropriate analysis can significantly reduce instrumentation cost and complexity. Fruitful applications of symbolic methods are thus favored in circumstances where robustness to noise, speed, and/or cost are paramount.

Using symbolic discretization as a data transform, although seemingly counter-intuitive, also has foundations in information and dynamics theory. For example, properties of symbolic encodings are central to the theory of communication[2], Markov chains for discrete systems[3], and bioinformatics[4]. (The interested reader is referred to Kitchens[5], which includes brief historical summaries at the end of each chapter.) These fields, in turn, have deeper roots, and have grown out of theories of language and games of chance. The objects of study in those fields (code words, DNA base pairs, and coin flips) are most naturally modeled as discrete states based directly on their physical attributes, hence a "symbolic" theory is obviously called for. It is not obvious, however, that a symbolic approach is useful when dealing with systems having continuous state spaces. Yet, as we show, many researchers have used symbolic transformation of continuous data with great success.

Symbolic treatment of time-series data is also closely related to the mathematical discipline of symbolic dynamics. The earliest developments in symbolic dynamics began with the study of the complex behavior of dynamical systems. In 1898, Hadamard developed a symbolic description of sequences in geodesic flows on surfaces of negative curvature.[6] Specifically, he identified a finite set of forbidden symbol pairs (those which cannot occur) and noted that possible sequences were those which did not contain the forbidden pairs. This work was extended by Morse[7] and later Morse and Hedlund[8], who examined

dynamical features such as periodic orbits in classical systems using a symbolic description. Morse and Hedlund were the first to use the term *symbolic dynamics*. Later, Collet and Eckmann[9] formalized symbolic dynamics by showing that a complete description of a dynamical system's behavior can be captured in symbolic dynamics.

Contemporary with Hadamard, Poincaré, in his 1899 analysis of the classic three-body problem, proposed that the complex time evolution of this system could be depicted using a kind of stroboscopic sampling of the multi-dimensional phase-space trajectory (see Holmes[10]). (Diacu and Holmes[11] provide an excellent recent summary of the development of symbolic dynamics in the analysis of celestial mechanics.) Specifically, Poincaré defined a surface in phase space, called a *surface of section*, such that the temporal evolution induces successive intersections between this surface and the (higher-dimensional) phase-space trajectory. The effect of this technique was to reduce the dimensionality of the problem and convert the continuous flow in phase space to a smooth discrete-time mapping between successive locations in the surface.

In a natural extension of this idea, other investigators found it useful to coarse-grain the Poincaré surface of section such that any intersections falling within a certain sub-region of the surface (called a *cell*) would all be designated with the same symbol. Following an orbit, the relative frequencies for intersections in various regions could then be statistically quantified, and the resulting temporal sequence of symbols could be studied as a replacement for the original variables (see Fig. 1.)
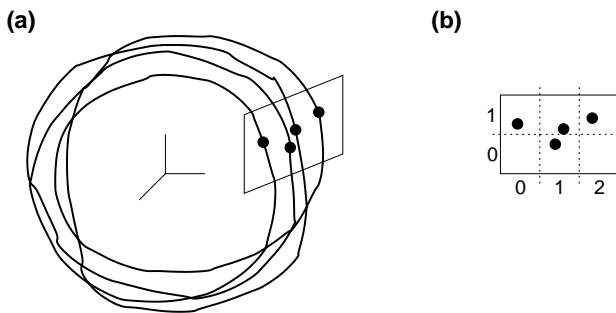

FIG. 1: Relationship of a Poincaré surface of section and a phase-space trajectory (a) with corresponding discretized section (b). The solid circles represent the points of intersection with the Poincaré sectioning plane.

Now consider the discrete-time evolution of an ensemble of points in the Poincaré surface of section. These points are assumed to obey identical deterministic dynamics and to differ only in their initial conditions. It is natural to ask about the time evolution of *ensemble averages* of quantities of interest (such as the average position, momentum or energy). We can also ask whether such ensemble averages are related in any way to *time averages* carried out following a 'typical' orbit. The answer to this question is of great practical interest because, typically, the ensemble averages are easier to calculate theoretically while the time averages of a few orbits are often the only quantities available to the experimenter. The relationship between these two very different approaches to the statistical description of a dynamical system is the central problem in *ergodic theory*.

It is possible to show that, provided there exists a probability density that is invariant under the discrete-time dynamics, the coarse-grained versions for ensembles are finite-state Markov systems. Ulam[12] conjectured that successive refinement of the coarse-graining would provide a convergent sequence of approximations to the (highly non-trivial) statisical evolution of the continuum behavior, desribed by the Frobenius-Perron operator.[13] More recently, Rechester and White[14,15] and Nicolis[16] have suggested that a refinement strategy based upon the dynamics could have improved convergence properties.

The developments in symbolic-dynamics have taken several diverse directions, including analysis of nonlinear oscillators, nonlinear maps, and connections to information theory and the notion of metric entropy. A discussion of the modern origins of symbolic dynamics and its relation to other fields is given in Jackson.[17] See also the recent collection of articles in Bedford *et al.*[18] for a sampling of related issues in pure mathematics. A general feature of this modern work in symbolic dynamics is that it is theoretical in nature, and most investigations rely on the existence of so-called generating partitions. Generating partitions divide the Poincaré plane such that each unique trajectory in phase space is associated with a unique sequence of symbols. This uniqueness requirement is particularly important for deterministic dynamics, where each initial condition produces a unique subsequent trajectory. It has been demonstrated that generating partitions can be constructed for certain classes of model systems, but there is no general approach for constructing generating partitions *a priori* when one is observing the behavior of an unknown system. In addition, it is also clear that generating partitions do not exist in the presence of experimental noise, even for systems with well-understood dynamics.[19] Thus, while symbolic dynamics provides a useful starting point for considering analysis of experimental time-series measurements, the theory is not sufficient for dealing with important practical concerns for experimentalists. To clarify this distinction between theory and application, we use the terms *symbolization* and *symbolic time-series analysis* to refer to our main subject here.

Because of the limitations of symbolic-dynamics theory, practical uses of symbolization have tended to be heuristic and empirical. In particular, it has been clearly demonstrated that heuristically defined symbolic partitions can be useful for characterizing temporal patterns without being generating. The first practical applications were probably associated with the advent of digital computing, where discretization was unavoidable. For early digital machines, such as those used by the British in the 1940s for air-warfare computations,[20] it was

observed that correlations among time-series measurements could still be accurately computed, even though these machines had only 7-bit precision. In more recent years, explicit applications of symbolization have proliferated far beyond digital computing to include such wide-ranging fields as astrophysics, classical mechanics, psychology and medicine, plasma physics, robotics, communication, linguistics, combustion, and multiphase flow.

In subsequent sections, we discuss common methods for constructing symbolic partitions, symbol trees, and measures of temporal structure and information content in the resulting symbol sequences. We also relate symbolization to other techniques that have been developed for the analysis of data from nonlinear and chaotic processes. Following that, we summarize how symbolic methods have been adapted to a variety of experimental contexts and needs. Finally, we end with a summary of current limitations in symbolic techniques and major issues in recent research.

II. PRACTICAL MEASUREMENT ISSUES

Experimental time-series measurements are typically acquired either at a fixed rate (per unit time) or synchronized to some triggering event. Fixed sampling rates are generally used when the dynamical process being observed is inherently time-continuous (that is, when we can assume that the dynamics are best modeled as the evolution of differential equations). Time-discrete measurements apply to dynamical processes that possess an inherent cycle and are more naturally modeled as maps (for example, population fluctuations between animal generations or changes between rotational cycles in machinery). As discussed above, one can convert time-continuous measurements to time-discrete measurements through the construction of a Poincaré surface of section (or Poincaré section for short). Poincaré sections are produced by continuously monitoring the system state and recording key variables when a specified triggering condition is met (e.g., when a hyper-plane in phase space is intersected in a particular way by the system trajectory). Such monitoring can be done in real time or, perhaps more frequently, during post-processing of previously recorded data. However the section is constructed, the resulting sequence of recorded values is discrete in time. In many instances, experimentalists only record a single observable at the moment of the triggering event. Depending on the dimensionality of the dynamics being observed, this often produces a lower-dimensional projection of the actual Poincaré section. An alternative procedure is to record the time intervals between successive occurrences of the triggering condition. Plots of successive pairs of observations or time intervals are often referred to as first return maps. Moon[21] provides a good overview of approaches used for triggered measurements.

Whether recording time-continuous or time-discrete data, it is important to properly account for the effects of aliasing by application of an appropriate anti-aliasing filter. Aliasing occurs whenever a dynamic process is observed using sampling rates that are too slow. The effect is to create apparent dynamical features that are artifacts of the sampling process rather than true characteristics of the original behavior. The basis for constructing anti-aliasing filters is the Shannon sampling theorem, which dictates that the sampling rate be at least twice as fast as any of the dynamics being observed. Steiglitz[22] and Oppenheim et al.[23] give good discussions of the effects of aliasing and how it can be avoided with the use of appropriate anti-aliasing filters.

Noise is another key issue in analyzing experimental measurements. The term noise can refer to measurement errors associated with the sensing device (measurement noise) or fluctuations in the dynamic state caused by external inputs (dynamic noise). In either case, it is implied that the noisy component is different from the processes of interest and of less importance to the observer. Frequently, this noisy component represents the effects of many independent or loosely coupled processes (i.e., the noise is high-dimensional), while the dynamics of interest are dominated by a few features and are low-dimensional. Ideally, low-dimensional dynamics will appear as distinct from the noise (e.g., large-amplitude, low-frequency variations versus small-amplitude, high-frequency variations). Unfortunately, this ideal situation is frequently not met in experimental practice, and the dynamics of interest can be mixed with undesirable features.

Noise can have important interactions with the symbolization process that can enhance or distort information content. For example, Cuéllar and Binder[24] found that for certain data adding a small amount of uncorrelated noise before discretization improved the effectiveness of noise-reduction techniques applied after discretization. In a different study, beim Graben[25] showed that symbolization can directly enhance signal-to-noise ratios.

In some cases, it is possible to make multiple measurements of dynamic systems over time. When all of the key variables are accessible, one can completely resolve the dynamical evolution of the system by means of simultaneously plotting all of the key variables, and the result is a direct reconstruction of the system phase space. In most cases, such complete observations are not possible, and the experimentalist must make do with a limited subset of the possible measurements. The most common situation is when only a single observable is available. Fortunately, time-delay embedding offers some hope for recovering at least some details of the unobserved variables (for good discussions on time-delay embedding, see Abarbanel[26] and Kantz and Schreiber[27]). The central objective of time-delay embedding is to reconstruct a facsimile of the phase-space dynamics of some multi-dimensional system from the observations of a single observable $\vec{X} = \{x(1), x(2), \ldots, x(N)\}$ by plotting the observations in a phase space of lagged coordinates

$\vec{\xi}(t) = \{x(t), x(t+\tau), x(t+2\tau), \ldots, x(t+(m-1)\tau)\}$. (See Fig. 2.) The variable τ is an embedding delay, and m is the embedding dimension. This approach for reconstructing multi-dimensional dynamics from a single observable relies on the topological equivalence of such a reconstructed trajectory as long as the reconstruction has dimension $m \geq 2D + 1$, where D is the dimension of the original system. Packard et al.[28] and Takens[29] both suggested that time-delay embedding could provide a way to extend limited experimental measurements for complex systems. Takens proved that, as long as the embedding dimension is sufficiently large, the reconstructed phase space is a true diffeomorphism of the original phase space. The underlying concept was further generalized by Sauer et al[30]. Since that time, time-delay embedding has become a widely used tool for data analysis. As we discuss later, there are some similarities between time-delay embedding and symbol-sequence analysis.
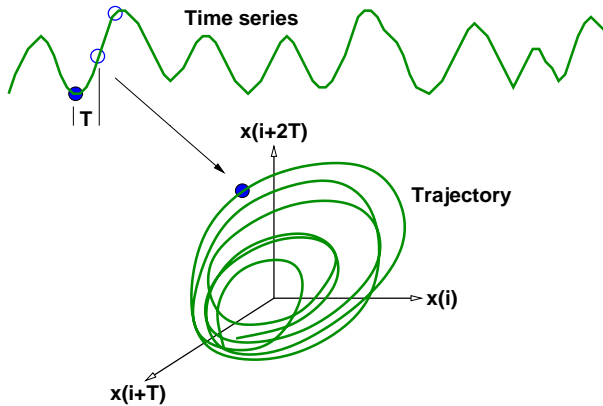

FIG. 2: Illustration of time-delay embedding.

Another important issue for experimentalists is that most statistical approaches for time-series analysis (with or without symbolization) assume that the observed process is at least locally stationary (that is, the system parameters are constant and external perturbations are minimal). If this is not true, then the statistical properties can vary over time and consistent comparisons become much more difficult. Nonstationarity is common for biological or large-scale natural systems such as the atmosphere or astrophysical objects. Another important example of nonstationarity occurs if the system under study is approaching a bifurcation (such as an instability threshold). The usual approach for dealing with potential nonstationarity is to try to limit the data-acquisition process to a relatively short period compared with any slow changes that the system may undergo. Analysis of repeated measurements can confirm whether stationarity was maintained. Methods such as those developed by Kennel and Mees[31] can also be applied to individual data sets to determine if shifts in the dynamics occurred during data acquisition. See also Kennel[32], Schreiber[33], Witt et al.[34], and Yu et al.[35] for alternative methods for testing for stationarity.

III. DEFINING SYMBOLS

Digital data recording automatically produces discretization (for example, 12- and 16-bit digitization are common). Such discretization, however, is generally much more refined than that used for symbolic analysis. An important caveat for experimentalists, however, is that the details of the digitization process itself can introduce confounding structure in the measurements that is unrelated to the process being measured. Specifically, Kapitaniak et al.[36] demonstrate that the behavior of certain analog-to-digital converters produces a nonlinear mapping that reduces measurement precision. They demonstrate the nature of these errors using symbolization and suggest methods for error reduction.

The most common approach for coarse symbol definition involves partitioning the range of the original observations (or the range of some transform of the original data such as the first differences between successive values) into a finite number of regions. Each region is associated with a specific symbolic value, and each original measurement is thus uniquely mapped to a particular symbol depending on the region in which the measurement falls. The number of possible symbols, n, is termed the *symbol-set size* (*alphabet size* in the symbolic-dynamics literature). For the simplest (binary) case there are two possible symbols and $n = 2$. In many cases, binary symbolization is convenient because it can directly exploit binary operations in computers. Hsu et al.[37] demonstrate that considerable improvement in computational efficiency can be produced by severe discretization, even when standard Fourier transforms are the objective. Higher values of n correspond to increasingly refined discrimination of measurement details, including the effects of any measurement noise that might be present. In the limit, when n equals the number of distinct values in the time series, the symbol series and original measured time series are equivalent in the sense that they contain the same information (that is, there is no longer any loss of information produced from the symbolization transform). Thus, in selecting the number of symbols, one inevitably chooses how much of the original information is retained in the subsequent analysis.

As noted previously, there is a theoretically optimal choice for locating partitions for noise-free, deterministic processes (see for example Crutchfield and Packard[19] and Crutchfield and Young[38]). Some general methods have even been proposed for estimating generating partitions for models (see Rechester and White[15], Grassberger and Kantz[39], and Davidchack et al.[40]). However, it is not possible to find generating partitions for most experimental observations because such partitions do not exist when noise is present, even in principle.[19] One is thus left with the practical problem of choosing appropriate partitions for a data set that may have been generated by an unknown dynamic process with unknown levels of noise.

Frequent *ad hoc* choices for the location of partitions between symbols are the data mean, midpoint or me-

dian, equal-size intervals over the data range, or regions of the range with equal probability (*equiprobable* or *equiquantal*[41] partitioning). For example, Tang *et al.*[42,43] found that a binary symbol set partitioned on the sample mean was quite adequate for reconstructing the dynamics of nonlinear models, even when the observed dynamics were heavily contaminated with noise. Rapp *et al.*[44] investigate errors of using midpoint, instead of median, partitioning and suggested a manner to check for spurious identifications of non-random structure. Hively *et al.*[45,46], on the other hand, used equal-sized data intervals to partition EEG signals in detecting precursors to seizures. Kim *et al.*[47] analyzed heart-rate dynamics using partitions aligned on the data mean and ±1 and ±2 sample standard deviations, for a symbol-set size of 6. Godelle and Letellier[48] used equiprobable symbols for analyzing measurements from free liquid jets in order to readily discriminate between random and non-random behavior.

In some cases, the context of the problem or the underlying physics dictates a natural choice for partitions. Systems involving dynamics with a natural threshold, for example, are a good case for physically based partitions. Specifically, neurobiological and chemical systems often exhibit an excitability threshold that must be exceeded for oscillations to begin (see Kádár *et al.*[49], Freund *et al.*[50], and Braun *et al.*[51], for example). When one is interested in observing the onset or absence of threshold crossings, the threshold value itself provides a reasonable choice for defining symbols. The presence of a limited number of distinctive dynamic states, such as the "stick" and "slip" condition in a dry friction oscillator (see Feeny and Moon[52]), also provides a natural partition choice.

However partitions are selected, sensitivity of the results to the choice of partition should be carefully evaluated. It is clearly possible to choose bad partition locations such that most, if not all, of the meaningful dynamical information is lost. Bollt *et al.*[53,54] illustrate this point forcefully in their investigation of the symbolic analysis of data from several different nonlinear models. Other investigators have attempted to systematize partition selection by iterating an initial set of partitions with an objective function reflecting the information content of the resulting symbolic series. Lehrman *et al.*[55] used such an approach for model data combined with a repeated assessment of the Shannon entropy for the symbol sequences (see the Symbol-Sequence Statistics section for a discussion of entropy). They reported that when enough symbols and appropriate partitions were used, entropy was maximized and the partition choice was "optimal". Godelle and Letellier[48] made similar arguments in their analysis of data from liquid jets, using examples from numerical models to confirm their assumptions.

Symbolization schemes based on first- or higher-order differences in observed measurements have also been proposed (see Kurths *et al.*[56]). These are effectively the same as range-partitioning schemes except that they operate on first or higher-order differences between sequential measurements in the original time series. Such schemes are sometimes preferred when the observed data are not fully stationary or where changes in time are more important than absolute measurement values. These transforms are termed *dynamic*, whereas those based on range-partitioning are termed *static*.[56] This dynamic, differenced-based symbolization has been employed by Bandt and Pompe[57] for their *permutation entropy* and in data mining and rule discovery (see Section VI G).

Still another approach for symbol definition involves partitioning the phase space rather than a scalar time series. Specifically, the symbols represent distinct regions of phase space or a subset of the phase space such as a Poincaré section. Observed symbol sequences, in turn, represent trajectory segments or mappings that link the separate regions according to the flow or mapping in phase space. Examples of this approach include Edwards *et al.*[58], who used a type of phase-space partition for analyzing the dynamics of model gene networks, and Hively *et al.*[45,46], who partitioned the time-delay reconstructed phase space (see next section) to identify seizure precursors in electroencephalograms. Baptista *et al.*[59] likewise used phase-space partitioning to model communication. Halow and Daw[60] labeled the transitions of trajectories between quadrants of reconstructed phase space to classify dynamics in fluidized-bed reactors. In contrast, Mischaikow *et al.*[61] and Lesher *et al.*[62] partitioned Poincaré maps to study the dynamics of a flexible ribbon and lamprey neural signals, respectively.

The concept of mapping flows in coarse-grained phase space is similar to the cell-to-cell mapping technique (see Hsu[63] and Tombuyses and Aldemir[64]).

IV. DEFINING SYMBOL SEQUENCES

After symbolization, the next step in identification of temporal patterns is the construction of *symbol sequences* (*words* in the symbolic-dynamics literature) from the symbol series by collecting groups of symbols together in temporal order. This sequencing process typically involves definition of a finite-length template that can be moved along the symbol series one step at a time, each step revealing a new sequence. If each possible sequence is represented in terms of a unique identifier, the end result will be a new time series often referred to as a *symbol-sequence series* (or *code series*[65]). Figure 3a illustrates this process for a time series that has been initially converted into a binary symbol series. In the example, the symbol sequences are constructed from the three successive binary symbol values occurring at each point in time. Each possible sequence is represented by its binary number equivalent (or the decimal value) determined by the position of each symbol in the template.

Symbol-sequence construction has at least outward similarities to time-delay embedding, and one might argue that the result of symbol-sequence construction is analogous to coarse-graining of the time-delay recon-
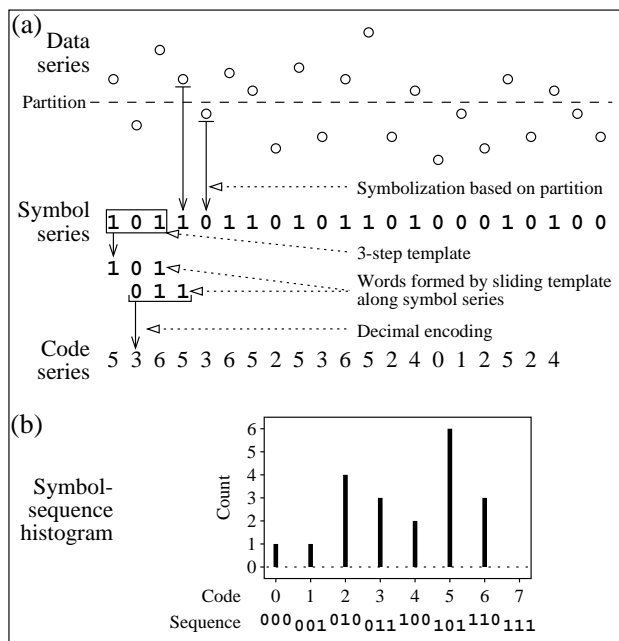
FIG. 3: Process of symbolizing a time series (a) and tabulating a symbol-sequence histogram (b).

of the length of individual branches depends on information theoretic measures that indicate how efficiently the observed dynamics are predicted (*i.e.*, how well the symbolization scheme compresses the available information). A similar approach is outined by Schürmann and Grassberger.[68]
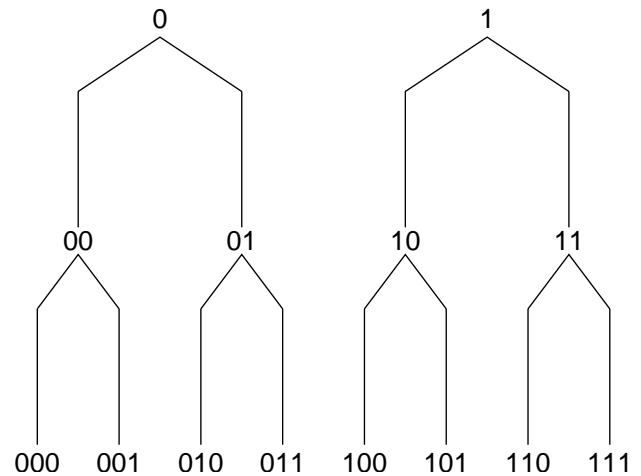

FIG. 4: Construction of a symbol tree.

structed phase space. Unfortunately, there is no rigorous analogue between the geometric construction underlying time-delay embedding and the informational content of symbol-sequence analysis, so there is not necessarily any finite symbol-sequence length that captures "all" of the available information in the sense of time-delay embedding. The time-delay embedding theorem promises that reconstruction will be faithful only for smooth transformations of the original phase space. For symbol sequences and symbol-sequence "space", the concepts of *neighborhood* and *continuity* can be retained[66], but *differentiability* is problematic.

Symbol-sequence construction has also been described in terms of symbol trees.[31,42] As illustrated in Fig. 4, the tree is composed of parallel branches, each of which represents a possible sequence of the available symbols. The length of the sequences determines the depth of the tree and thus the number of branches. For a fixed sequence length of L successive symbols, the total number of branches is n^L, and thus the number of possible sequences increases exponentially with tree depth. Many investigators have considered only patterns with a single fixed sequence length, thereby simultaneously including all the parallel branches in the tree. Such a constraint is convenient for enumerating the statistics of possible sequences, but it neglects realistic behavior in which some nonrandom patterns occur over longer intervals than others. A recent improvement to the fixed-sequence-length approach is the implementation of context trees by Kennel and Mees.[31,67] This approach allows some of the possible sequences (*i.e.*, branches) to be shortened to reflect reduced predictability over long times. Modification

For any given dynamical system, all sequences are not realizable. Such nonoccurring sequences are called *forbidden sequences* or *forbidden words*. Graphically, on a symbol tree, they represent branches which are trimmed such that longer sequences which contain a shorter, forbidden sequence cannot occur. D'Alessandro and Politi[69] discuss dynamical data complexity and its relation to forbidden sequences.

Closely related to the issue of symbol-sequence length is the question of data sampling rate. This is different from the Nyquist aliasing issue discussed earlier and is instead associated with the problem of symbolic redundancy. For discrete measurements, one normally samples the system at each natural iterate, and each step provides important new information about the system state. However, it is common for continuous experimental data to be measured at rates so high that the resulting symbol series contains multiple successive repetitions of the same symbol. From the standpoint of observing meaningful patterns, high frequencies of symbol repetition are not very useful and usually indicate over-sampling of the original data. On the other hand, if the important time scales in the measured signals are much shorter than the sampling interval, one is apt to create aliasing or lose information about the instability time scales (*i.e.*, maximum Lyapunov time scales). This problem is very similar to the issue of choosing an appropriate time interval for time-delay embedding, and thus similar approaches have been adopted for symbolic analysis. The usual approach for reducing symbol redundancy is to lengthen the inter-symbol time interval used in constructing symbol sequences and/or applying some type of downsampling to the original data. Commonly used tools for establish-

ing a reasonable downsampling and/or inter-symbol interval are the autocorrelation function[21] and the mutual-information function[70], which is defined by

$$I(\tau) = \sum p_{i,j}(\tau) \log_2 \frac{p_{i,j}(\tau)}{p_i p_j} \qquad (1)$$

where τ is a specified time delay between successive measurements. The partitioning used to determine the probabilities in the above equation is typically based on equiprobable binning of the observed data to avoid non-zero values of mutual information when measurements become truly independent. Fraser and Swinney[70] were the first to propose the use of mutual information as a tool for evaluating the time interval used for time-delay embedding. In this seminal work, an estimate of the mutual information of the original (*e.g.* continuous-time analog) signal is computed by successive refinements of the (discrete-time symbolic) partition until convergence is achieved. A direct comparison can then be made between the autocorrelation function of the continous signal and the mutual information.

It is also possible to introduce estimators of correlation which refer only to the symbolic form of the data. Such "symbolic" correlation estimators are of obvious importance when dealing with purely discrete phenomenon (as in DNA or linguistic analysis[71–73]). Another situation where symbolic correlation estimates are useful arises when one needs a fast estimate in real time[74], or when one is dealing with highly compressed data due to memory limitations.[37] More recently, Roulston[75] demonstrated procedures for estimating *a priori* the uncertainty in the mutual-information function.

It is of great interest to understand more fully the relationship between the symbolic correlation estimates, performed on coarse-grained versions of the signal, and the more traditional "autocorrelation" function, $C(\tau)$, which uses the analog form of the signal. The question is important because much is known about the effects of noise and linear filters upon $C(\tau)$ while very little is known about analogous effects upon the related symbolic signal. For example, it has been known for some time that if $x(t)$ is a stationary, zero-mean Gaussian linear process, there is a direct relationship between the zero-crossing probability measured at the sampling interval τ and the autocorrelation function $C(\tau)$. This relationship can be traced to properties of Gaussian integrals known since the late 1800s.[76] If we assign the symbol 0 to $x(t) \leq 0$ and 1 to $x(t) > 0$ (*i.e.* we *hard clip* the signal) then, using our present notation, the autocorrelation $C(\tau)$ and the probability that a comparison of $x(t)$ and $x(t+\tau)$ detects a bit flip, $p_{01}(\tau) + p_{10}(\tau)$, are related by[77] (as cited in Ref.[78], p. 57-58)

$$C(\tau) = cos(2\pi p_{01}(\tau)). \qquad (2)$$

Here, p_{xy} represents the probability of observing sequence xy. (*N.B.* we have replaced $p_{01}(\tau) + p_{10}(\tau)$ by $2p_{01}(\tau)$ in (2) by the following argument: once the signal flips the bit "up", $0 \to 1$, it must flip the bit back

"down", $1 \to 0$, before another $0 \to 1$ transition can be observed. Hence, for a long time series we will have $p_{01} = p_{10}$ to high accuracy even for non-Gaussian, non-stationary systems which remain zero-mean processes.)

The result (2) can be extended to continuous-time stationary zero-mean Gaussian linear processes and reduces to the formula of Rice relating the instantaneous zero-crossing rate with the second derivative of the autocorrelation at zero time delay.[76,79] This can be seen as follows: the instantaneous bit-flip 'rate', *i.e.* the number of bit flips per unit time, is defined to be

$$\lim_{\tau \to 0} \frac{2p_{01}(\tau)}{\tau} = 2\dot{p}_{01}(0). \qquad (3)$$

(Overdots denote differentiation with respect to τ.) Taking the limit as $\tau \to 0$ of (2) we find Rice's result (we assume the normalization $C(0) = 1$):

$$\dot{p}_{01}(0) = \frac{1}{\pi} \left[-\ddot{C}(0) \right]^{1/2} \qquad (4)$$

Kedem[76] discusses a generalization of (2) to three-point autocorrelation functions, and notes that there are no higher-order results known even for Gaussian linear processes. Kedem has extended (2) to some non-Gaussian processes, as discussed in Ref.[76], where necessary extensions of Rice's results are also discussed (see also[80]).

An attractive property of symbol-sequence statistics is that they provide a compact summary of multi-step correlations (even if their relationship to the more familiar multi-point linear correlation functions are not understood at this time). Tang *et al.*[81] have numerically examined the relationship between the linear autocorrelation, the mutual information as computed by Fraser and Swinney,[70] and the symbolic mutual information computed using (1) for a binary coarse-graining of analog signals. As shown in Fig. 5, the autocorrelation and mutual-information functions for purely periodic behavior follow repeating cyclical variations that match the natural period. For chaotic processes, however, these functions decay to a zero value after a finite time. For "random" processes, the correlation functions immediately tend to a near-zero finite value, indicating no correlation over all timescales. (In the figure, mutual information was normalized according to $\rho = \sqrt{1 - \exp(-2I)}$.[82]) When determining embedding delay or inter-symbol time intervals, one typically chooses some significant fraction of the time interval to the first zero in autocorrelation or first minimum in mutual information.[26] Another approach is to identify a natural sampling period based on the physics of the problem. For example, Godelle and Letellier[48] used a time interval that was a fraction of the driving frequency of vibrations applied to their liquid jet. As with symbolic partition definitions, it is always a good idea to determine how sensitive the final results are to the chosen inter-symbol interval.

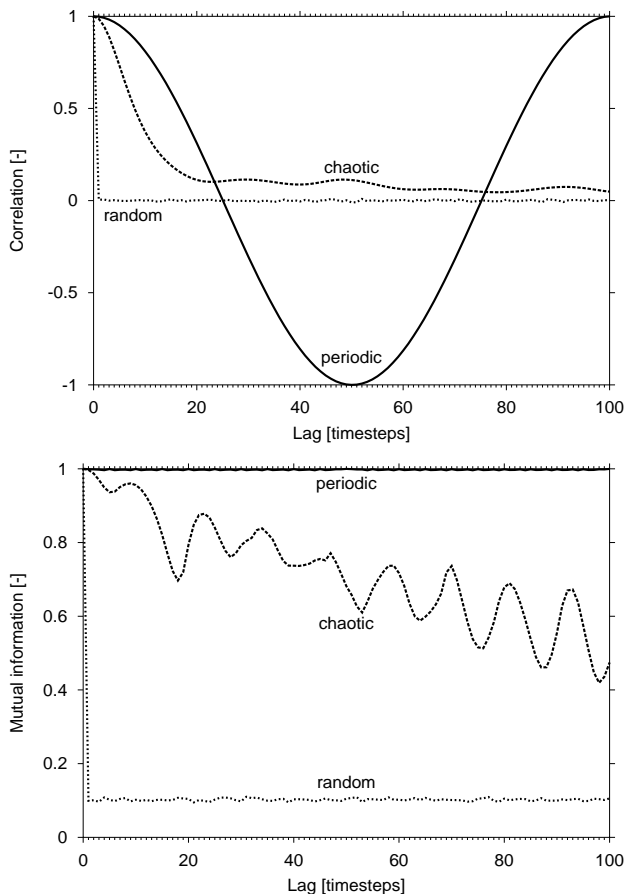In many experimental situations, it is possible to record multiple measurements either simultaneously or

FIG. 5: Typical behavior of the autocorrelation (a) and mutual-information (b) functions for periodic, chaotic, and random data.

in some regular time or phase relationship to one another. This is most often the case for spatially extended systems where there are multiple components or interacting regions. One thus obtains more than one time series, each of which can be symbolized individually or in combination with the others. One obvious approach is to combine individual symbols into multivariate symbol sequences (*i.e.*, symbolic vectors), where each signal is assigned a specific position in the symbolic word. Depending on how the data are originally recorded, the resulting words could represent a single instant in time or span some fixed period of the dynamics. Such a treatment of multivariate measurements is particularly useful for detecting and characterizing synchronization, in which groups of components assume some fixed dynamical relationship with each other (for example, in terms phase or amplitude). Moon *et al.*[83] illustrate the utility of this approach for analyzing the behavior of multiple coupled impact oscillators. Daw *et al.*[84] used a similar approach for studying interactions among multiple cylinders in internal combustion engines.

The mutual-information concept was recently extended by Schreiber[85] to the more general idea of transfer

entropy, which is appropriate for evaluating relationships among multiple components in extended systems (*e.g.*, identifying driving versus responding elements). Transfer entropy is defined by

$$T_{J \to I} = \sum p(i_{n+1}, i_n^{(k)}, j_n^{(l)}) \log_2 \frac{p(i_{n+1} \mid i_n^{(k)}, j_n^{(l)})}{p(i_{n+1} \mid i_n^{(k)})} \quad (5)$$

Transfer entropy can be directly estimated from symbolized (*i.e.*, partitioned) data, but Schreiber also proposes an alternative method based on the correlation integral.

For uniquely identifying symbol sequences, it is common to use letter strings, in which the letters represent symbols, or, alternatively, an index computed from numerical values assigned to each symbol. For the latter convention, numerical symbol values typically range from 0 to $n-1$, where n is the symbol-set size. A unique index for each possible sequence can then be determined by the base-n value of the sequence. The values of each successive symbol in the sequence are then weighted by n^{i-1}, where i is the relative position of the symbol in the sequence (in either forward time or reverse time order). For example, if $n=3$ the length-5 sequence $01020 \to 0 \times 3^4 + 1 \times 3^3 + 0 \times 3^2 + 2 \times 3^1 + 0 \times 3^0 = 33$ (base-10). The resulting sum for the sequence can then be expressed as a single number in any convenient base such as 2 or 10. Whichever method is used to designate symbol sequences, the symbolization and sequence-identification processes transform the original time-series data to a symbol-sequence series. The statistics of these sequence series is ultimately the essential focus of experimental data analysis.

V.   SYMBOL-SEQUENCE STATISTICS

Temporal structure in observed data is revealed by the relative frequency of each possible symbol sequence. Various types of statistics can be determined from the estimated symbol-sequence probability distribution. Direct visual observation of the frequencies with symbol-sequence histograms provides a convenient way for observing possible patterns. The usual histogram format depicts the sequence identification index (*e.g.*, as defined by the base-n method described above or a unique character string) on the abscissa versus the observed frequency for that index on the ordinate as illustrated in Fig. 3b. One important use for such plots is rapid detection of experimental or data-handling errors (*e.g.*, over-sampling or nonstationarity). Also, one can be quickly alerted to sudden shifts or bifurcations in the dynamics of an experiment as operating or observational parameters are changed. This type of application for visual histograms is illustrated by Godelle and Letellier[48] for their studies of free liquid jets.

Beyond visual inspection, symbol-sequence analysis depends on quantitative measures of symbol-sequence frequencies. Such measures can be divided into two general groups: those based on classical statistics (more or

less) and those based on information theory. Important examples of the former include the Euclidean norm and chi-square statistics, which are usually defined, respectively, as

$$T = \sqrt{\sum_i (X_i - Y_i)^2} \qquad (6)$$

and

$$\chi^2 = \frac{\sum_i (X_i - Y_i)^2}{\sum_i (X_i + Y_i)} \qquad (7)$$

A principal use for the above type of statistics is to quantify the difference between two symbol-sequence histograms. Tang et al.[42,43] were the first to employ the Euclidean norm statistic as a symbolic objective function for fitting nonlinear model parameters based on noisy experimental observations. The chi-square has been recently used by Kennel and Mees[31] for evaluating stationarity. As Kennel and Mees point out, care must be used in applying the usual chi-square statistical confidence intervals for testing hypotheses because these intervals are based on an assumption that the frequency of each sequence is independent of the frequencies of other sequences. Such independence is typically not achieved for correlated time-series data. In addition, the method of symbol-sequence construction can result in a redundancy artifact between the frequencies of different sequences that reduces the actual degrees of freedom. A key objective of the "weighted context tree" approach proposed by Kennel and Mees[67] is to minimize such redundancy.

Examples of information-theoretic measures for symbol-sequence frequencies include the Shannon and order-q Rényi entropies defined, respectively, as

$$H = -\sum_i p_i \log_2 p_i \qquad (8)$$

and

$$H^q = \frac{1}{1-q} \log_2 \sum_i p_i^q \qquad (9)$$

where p is the histogram of symbol-sequence frequencies. The base-2 logarithm places the entropies in units of bits.

An important use for these and similar measures is to evaluate the relative complexity of the symbol-sequence frequencies. Specifically, broad symbol-sequence frequency distributions produce high entropy values, indicating a low degree of deterministic structure. Conversely, when certain sequences exhibit high frequencies, low entropy values are produced, indicating a high degree of determinism (low entropy is also a characteristic of over-sampled data). The theoretical connections between entropy and noisy nonlinear systems were first fully explained by Crutchfield and Packard.[19] Good discussions of complexity measures related to the above are

given by Kurths et al.,[56] Pincus,[86] Kennel and Mees,[67] Perry and Binder,[87] and Rapp et al.[88]

Finite-sample effects have been shown to significantly affect entropy estimates.[89–91] Specifically, for increasingly longer sequences from a finite-length time series, entropy tends to be underestimated. Herzel[89] offered the expected value of entropy for length-L sequences:

$$\langle H_L \rangle \approx \sum_{i=1}^{M} -p_i \log p_i - \frac{M}{2N} \qquad (10)$$

where M is the number of sequences with $p_i > 0$ and time series of length N. These results were followed upon in the work of Hertzel, Schmitt and Ebeling[90] and an improved correcting formula by Grassberger[91] based on higher moments was discussed:

$$H_L^{grass} = \sum_i \frac{L_i}{N} \left(\log N - \Psi(L_i) - \frac{(-1)^L}{L_i + 1} \right) \qquad (11)$$

for $\Psi(x) = \mathrm{d} \log \Gamma(x)/\mathrm{d}x$.

An interesting topic of symbolic descriptions of complicated time series is measures of complexity. The Shannon and Rényi entropies are two such measures, but they do not necessarily approximate the true, dynamical entropy of the source which generated the time series. One such estimator is the Effective Measure of Complexity given by Grassberger.[92] This is a measure relating residual or truncated entropy and dynamical-entropy (i.e. Kolmogorov-Sinai) estimates. For truncated entropy

$$h_L = H_{L+1} - H_L \qquad (12)$$

and Shannon entropy per sequence step

$$h = \lim_{L \to \infty} h_L , \qquad (13)$$

then the EMC is

$$\mathrm{EMC} = \sum_{L=1}^{\infty} (h_L - h) . \qquad (14)$$

The EMC is a lower bound for the true measure of complexity.

Crutchfield and Young[38] used a statistical-mechanical approach to describe information-processing complexity. The ϵ-machine is a computational model constructed from a data set which attempts to describe in a minimal way the data patterns. The measure of complexity was defined to be a form of Rényi entropy.

In many cases, the ultimate objective in generating symbol-sequence statistics is to test null hypotheses about the observed data. Because there are uncertainties in determining appropriate confidence limits for many time-series statistics, researchers often rely on the use of surrogate data for bootstrapping the expected distributions under the null hypothesis. Good discussions of the generation and use of surrogate data are given by

Theiler *et al.* ,[93] Theiler and Prichard,[94] Schreiber and Schmitz,[95] and Dolan *et al.* .[96] Regardless of the specific generation method, the general approach is based on creating many realizations of time series (surrogates) that are consistent with some specific null hypothesis and, in every other way, also consistent with the observed data.

A simple example of surrogate generation is to repeatedly randomize the time order of the observed data to create many examples of the same measurements with exactly the same measurement frequency distribution but with any temporal structure removed (for example, see Chorafas[97]). One can then establish the confidence limits for evaluating some test statistic from the original data against the random null hypothesis by repeatedly generating the test statistic for many examples of the random surrogates. An instance of this is the computation of a "Monte Carlo" probability by Rapp *et al.*[44] In another instance, Schwarz *et al.* use Monte Carlo probability to estimate the error in mutual-information functions.[98] More sophisticated surrogates are required for testing other hypotheses for the generating process (*e.g.*, the shuffled Fourier-transform methods described by Schreiber and Schmitz[95]). Surrogates can also be generated for symbolic data, either by applying standard surrogate-generating techniques to the original time-series data before symbolization, or by using the symbolized data directly. Van der Heyden *et al.* ,[99] for example, demonstrated a procedure for making symbolic surrogates to test the null hypothesis that the observed data are consistent with an *n*th-order Markov generator.

One of the most recent uses for symbols in analyzing data from unknown sources is to test the hypothesis that the observed data could have been generated by a Gaussian linear process (or at most filtered by a static nonlinear transformation). In the context of nonlinear dynamics, if one observes a significant amount of time asymmetry, the above hypothesis can be clearly rejected.[100–102] In many cases, this rejection is extended to indicate that there is a strong possibility that the generating process was, in fact, nonlinear. The key relevant feature in this context is the presence of significant time asymmetry, that is, a difference in the symbol-sequence statistics depending on whether one observes the data in forward or reverse time. Voss and Kurths[103] and Daw *et al.*[104] have proposed two different symbol-based methods for testing the linear Gaussian hypothesis.

VI.   APPLICATIONS

In this section we summarize references from several different disciplines in which symbolization has been successfully applied. Our goal is not to create an exhaustive list, but rather to provide a sufficient breadth of examples such that most readers will be able to find problem contexts that they can relate to. In the end, one of the most important observations from previous work is that the theoretical foundation of symbolic analysis is still in its infancy. We expect that the greatest near-term benefits from its use will probably come from individual modelers and experimenters adapting previously demonstrated approaches to their specific needs and interests.

A.   Astrophysics/geophysics

An early application of symbolization to astronomy was made by Goldstein[74], who used hard clipping to analyze weak reflected radar signals from the planet Venus to measure the rotational period. Given the original signal resolution, it was estimated that several thousand hours of computer time (on existing computers) would have been needed to measure the line shape using Fast Fourier transforms. Single-bit digitization allowed the construction of a special-purpose computer to calculate the correlation function in hardware in real time as the signal arrived, and allowed the lineshape to be measured in only 1.4 hours.

More recent applications of symbolic analysis to astrophysics and geophysics have centered mostly on the interpretation of complex signals arising from earth-based observations of complex astrophysical and geophysical phenomena. In the latter case, Gavrishchaka and Ganguli[105] noted that threshold-based symbolization of measurements of auroral electrojet dynamics permitted improved forecasting of large-amplitude events with neural networks. The authors attribute this improvement to the fact that the symbolization minimizes the impact of small-amplitude details in the measurement signals that are not related to the dynamics that dominate the large-scale events. The authors also suggest that, in practice, it may be advisable to set up and train multiple neural networks (*e.g.*, in parallel) with varying threshold levels so that determination of the optimal threshold level and training at that level can be accomplished simultaneously.

Schwarz *et al.*[98] used a binary symbolization to analyze solar flare events with mutual information, Shannon entropy, and an algorithmic complexity measure. They found that, based on the characteristics of the symbolic measures, the flare spikes were more likely caused by local organization (a single event exciting many adjoining areas) rather than global organization (a succession of events in the same area).

B.   Biology and medicine

There have been many recent applications of symbolic analysis for biological systems, most notably for laboratory measurements of neural systems and clinical diagnosis of neural pathologies. Lesher *et al.* ,[62] for example, symbolized experimental time series data from lamprey locomotion. They visualized spike time series from bursting oscillators in the spinal cord as Poincaré sections derived from embedding the fast (spike) oscillations

in a higher-dimensional phase space. Symbols were produced from regions on the Poincaré map. Freund et al.[50] and Greenwood et al.[106] studied the electrosensory response of paddlefish to noise and electric signatures from individual Daphnia plankton. They employed symbolic models and detection of phase synchronization between the receptors and external noise to demonstrate the importance of stochastic resonance in the paddlefish's ability to detect their prey. The symbolic partition in this case was defined by the characteristic detection threshold of the paddlefish's electroreceptor cells. Steuer et al.[107] examined interspike-interval sequences of neurons from paddlefish and crayfish using techniques from symbolic dynamics and information theory. They symbolized the time intervals between firings of neuronal elements and analyzed the symbolized temporal sequences with a measure of local predictability. With this analysis, they noted the differences between the two types of neural responses associated with crayfish's ability to "recognize" patterns simulating prey activity.

Wendling et al.[108], bein Graben et al.[109] and Hively et al.[45,46] all used various implementations of symbolic analysis for characterizing electroencephalogram (EEG) signals. Wendling et al.[108] focused on stereoelectroencephalographic (SEEG) signals recorded with depth electrodes to understand interactions between brain structures during seizures. They proposed a comprehensive methodology for comparing SEEG seizure recordings that involves three key steps: segmentation of SEEG signals; characterization and labeling of segments; and comparison of observations coded as sequences of symbol vectors. beim Graben et al.[109] applied symbolic analysis to nonstationary and noisy multivariate EEG signals in order to estimate event-related potentials (ERP). Their approach included cutting the continuous time-serial data into epochs according to the stimuli events presented to the subjects. They then employed a statistical mechanics approach to coarse-grained symbolic descriptions of the dynamics and developed time-dependent measures of complexity that could be monitored to detect changes associated with the stimuli. Their findings indicated that symbolization could be useful for investigating synchronization and phase locking of neuronal oscillators in the context of ERP studies. Hively et al.[45,46] developed a method for diagnosing EEG changes based on shifts in the phase-space density functions as represented by a discrete coarse-graining of phase space. In the approach reported here, the criterion for identifying individual bins (i.e., phase-space symbols) was selected based on equal-interval partitioning of the observed signal range. Euclidean and chi-square norms were used to detect shifts in the phase space densities associated with epileptic seizure precursors.

Saparin et al.[110] used symbolic techniques to encode two-dimensional images of human cancellous bone and analyze the spatial complexity as a function of structural changes due to osteoporosis. The symbolic transform was based on detection of both absolute intensity and edge features in 2D CT scanner images. Taken together these features defined an encoding set of five symbols. Using measures of complexity determined from the resulting symbols, the authors determined that the bone complexity declined more rapidly than density with the loss of bone due to osteoporosis.

Edwards et al.[58] studied a simple ODE model for a genetic network in which model genes deterministically control the production rates of other genes. The dynamics of these equations are found to be represented symbolically, with symbols defined on the basis of the flow of the trajectory through the state space and the intersection of trajectory with specified boundaries. The resulting symbol letters and words correspond to Poincaré maps of the integrated flow. With this information, the authors discuss the solution of the reverse problem of determining the underlying network from the observed dynamics.

Kurths et al.[111] considered both static and dynamic symbolic transformations to ECG signals to characterize heart-rate variability. They analyzed the RR intervals, and first difference of the intervals, of the cardiac cycle using both the Shannon and Rényi entropies as measures of signal complexity. With a symbol-set size of 4 and sequence length of 3, they found that the ECGs of persons with cardiac risk exhibited more ordered symbolic patterns than those without; the generalized Rényi entropy was found to be more useful than Shannon entropy. The authors found that combining information from the symbolic analysis with traditional frequency-domain (Fourier) analysis yielded better results than the symbolic or Fourier analysis alone.

Kurths et al.[56] also applied symbolic methods to cognitive psychology, particularly regarding synchronization of keystrokes. Using a unique symbolic encoding, they were able to measure phase shifts between left- and right-hand keystrokes to polyrythms on an electronic keyboard as a function of driving tempo (see Fig. 6). In the figure, the left hand was keying 3 strokes for every 4 strokes in the right hand, and the figure maps the loss of coordination with driving tempo.

FIG. 6: Symbolic representation of 36 left-hand keystroke intervals as a function of driving tempo. Keystrokes were performed on an electronic keyboard with 3 left-hand strokes per 4 right-hand strokes. Black pixels represent interstroke intervals longer than their immediate predecessor, white else. From Kurths et al. (1996).[56]

C. Fluid flow

Application of symbolization to fluid flow measurements has spanned a wide range of data types from

global measurements of flow and pressure drop, to formation and coalescence of bubbles and drops, to spatiotemporal measurements of turbulence. Rao and Jain[112] developed an approach for transforming images of complex flow fields (as well as other textured fields) into a symbolic representation. Symbol sets were defined based on phase portraits reconstructed from flow field images. After the symbolic transformation, they used symbolic models to reconstruct salient features of the original image (i.e., they demonstrated symbolic compression of the images).

Lehrman et al.[55] applied symbolic analysis to the Lorenz ODE model, which is related to fluid velocity and temperature fluctuations in Bérnard thermal convection. Their objective was to demonstrate characterization of the dynamic coupling between different time signals coming from the same process (i.e., demonstrate that the signals did indeed come from the same process). In this case, the different signals were different variables in the Lorenz model. They also used the same technique to quantify coupling between different components in high-dimensional dynamics produced by a system of multiple coupled nonlinear equations proposed by Lorenz[113] for simulating the key features in atmospheric turbulence. A version of this method was used previously by Mazzucato et al.[114] and Rechester et al.[115] to analyze turbulent plasma measurements.

Lehrman and Rechester[116] developed a method for extracting symbolic cycles in time series from turbulent flow systems. Using a dynamic partioning, they developed a stability factor to describe the cycles. They applied their analysis to a water flow in a pipe with $D = 0.3$ m and $L = 30$ m at $\mathrm{Re_D} = 3 \times 10^5$. An interesting result was that although time records from wall and interior measurements had no direct correlation, they had similar symbolic cycle distributions, suggesting the dynamic patterns at the different measurement locations were similar.

Godelle and Letellier[48] employed symbolic methods for evaluating experimental measurements from free liquid jets of water and water-glycerol mixtures. Time series measurements of jet diameter were made using illumination from a laser sheet located at varying distances below the injector. The authors constructed first return maps and generated symbols from the diameter measurements using an equiprobable partition of the range. Special care was taken to evaluate the effects of changing both the number of available symbols and the sequence length. Extensive use was made of symbol sequence histograms in the evaluation process. On the basis of their results, Godelle and Letellier concluded that the jet dynamics ranged from white noise to various types of deterministic intermittencies. See Fig. 7 for symbol statistics of a range of nozzle-orifice sizes.

Angeli et al.[117] used a symbolic transform developed by van der Welle[118] to identify bubble passages in three-phase flow. The signal from a laser and light sensor was transformed into a binary series, a 1 for a bubble in the

FIG. 7: Symbol statistics at a constant jet velocity over a range of orifice diameters. From Godelle and Letellier (2000).[48]

light beam and 0 else, and the binary series was analyzed statistically.

Gonçalves et al.[119] examined the interdrop-interval time series from a dripping faucet experiment. Using a topological-entropy metric to define partitions, they constructed minimal topologcal machines and developed a new graphical method to describe the dynamics. They found that the dynamics from three distinct flow regimes could be described with the same basic topological graph with some differences in weightings of certain graph branches.

D.   Chemistry

Chemistry-related applications of symbolic techniques have been developed for chemical systems involving spontaneous oscillations or propagating reaction fronts. An example of the latter is given by Jung et al.,[120] who studied the formation of coherent structures in model 2D reacting systems. In their analysis of spatiotemporal patterns produced in a cellular excitable medium and a reaction diffusion model of CO oxidation, Jung et al. developed a method for defining the complexity of coherent reaction clusters (i.e., distinctive regions where significant reactions are occurring). These clusters evolve over time by collision and subsequent merging to produce a distribution of cluster sizes. Although they did not explicitly use the term symbolization, the authors developed a procedure for discretizing the observed patterns into binary values based on a threshold of activity. Active sites were then further grouped into clusters depending on their proximity as immediate neighbors. The degree of homogeneity in the distribution of clusters was then evaluated using a kind of spatiotemporal entropy. It was found that this entropy was a useful measure of cluster pattern changes as model parameters were changed.

Hsu et al.[37] applied a type of symbolization for improving the performance of Fourier-transform ion-cyclotron mass spectrometry. Specifically, they used 1-bit discretization of the original signals to greatly reduce data storage, search, and retrieval requirements (the storage requirements alone were reduced by a factor of at least 20). At the same time, they demonstrated that the resulting mass spectra were as useful and, in some cases, enhanced over the original spectra (see Fig. 8 for an example comparing mass spectra of high- and low-precision signals). The major change required in order to use the high level of discretization was to construct a new com-

pound library with a similar degree of discretization.

FIG. 8: Comparison of mass spectra from raw time series (a) and its 1-bit representation (c). From Hsu (1985).[37]

E.   Mechanical systems

Mechanical systems were one of the first applications where symbolic analysis was successfully used to characterize complex dynamics. Feeny and Moon[52] were able to demonstrate good agreement between experiment and a simple model using a symbolic description of the dynamics of a dry friction (stick-slip) oscillator. As described previously, this was a case where the binary symbol selection was easily defined from the naturally discrete state of the system (sticking or not sticking). In addition to explaining the observed dynamics, Feeny and Moon demonstrated the computation of a symbolic autocorrelation function and used it to estimate the largest Lyapunov exponent.

In another paper, Moon et al.[83] demonstrated the use of symbols for describing the spatially complex, temporally chaotic dynamics of eight coupled impact oscillators connected by a string. Their experiments utilized triggered, multi-channel recordings of the impacts of each oscillator as the string was vibrated with an electromagnetic shaker. Symbols of 0 or 1 were generated directly by the detection system during each drive cycle according to whether an impact had occurred or not. The global system state for each drive cycle was then represented as an 8-bit binary number as a function of time. Note that because the original 8-bit state vectors were directly recorded, this is an example of direct detection of the symbols as opposed to post-transformation of the original data to symbols. A symbolic entropy measure was used to characterize the dynamics as operating parameters were changed (see Fig. 9). Moon[121] also gave a comprehensive overview of his group's experimental techniques for measuring chaotic behavior in mechanical systems that included a brief discussion of symbolic analysis.

FIG. 9: Symbol entropy of an impact oscillator as a function of constraint gap. From Moon et al. (1991).[83]

More recent studies by Kobes et al.[122] studied the behavior of a 3-ODE model for a type of periodically driven pendulum. The specific model they studied corresponds to a physical system consisting of two gears and a rod. The authors used characteristic regions of a Poincaré section to define a 3-letter symbol dynamics, which they used to characterize the behavior of the system as parameters were changed. Through their numerical experiments, Kobes et al. determined similarities between this system and other well-known model systems, including the forced Brusselator and the dissipative standard map.

Daw et al.[65,84,104] applied symbolic methods to the analysis of experimental combustion data from internal combustion engines. Their objective was to study the onset of combustion instabilities as the fueling mixture was leaned. In-cylinder pressure measurements were converted to discrete time series of heat release for each cylinder and each engine cycle. The symbolization procedure involved equiprobable partitioning in a fashion similar to that used by Godelle and Letellier[48] for liquid jets. Also like Godelle and Letellier, the authors evaluated the effect of changing the symbolization parameters on the symbol-sequence histograms and complexity measures. In addition, they utilized measures of time asymmetry to observe details of the bifurcation sequence underlying the combustion instability.

F.   Artificial Intelligence, Control, and Communication

In a prescient early paper, Kalman[123] examines the control of nonlinear systems given sampled data. He discusses the interplay of sampling and nonlinearity, the extreme sensitivity to initial conditions of many nonlinear systems, the statistical nature of coarse-grained (symbolic) representations of the sampled data, and the relationship to finite-state Markov models. More recently, Delchamps[124] has examined strategies for stabilizing linearly unstable systems if one is given only coarse-grained information about the state of the system.

An example application of symbolization to communication is the study by Dolnik and Bollt[125,126], who used small perturbations to encode messages in oscillations of the Belousov-Zhabotinsky (BZ) reaction. Binary symbolic messages were encoded by forcing the chaotic oscillations to follow a specified trajectory (see Fig. 10). In addition to demonstrating communication with the BZ system, the authors' objective was to study several practical aspects of applying symbolic analysis in a real-world noisy laboratory environment. Practical issues investigated include modeling noisy time series, learning the underlying symbol dynamics, and evaluating system response to parametric changes.

Data compression is currently of great concern for increasing effective digital communication bandwidth. Goodman and Brooke[127] developed a symbolic substitution system for data compression by mapping symbol strings onto a two-dimensional tree. The resulting two-dimensional pairs could be "transmitted" and then subsequently mapped to another tree structure for regener-

FIG. 10: Control of the Belousov-Zhabotinsky reaction to communicate the message "Chaos". From Dolnik and Bollt (1998).[126]

ating strings at the "receiving" end. Both adapative and nonadaptive versions of the system were defined.

Tani[128] describes an artificial intelligence application in which a robot constructs a symbolic model that guides its interactions with the environment. He focuses on two essential problems for applying symbols in artificial intelligence. The first problem is symbol grounding, which is a generalization of how symbols are defined so that they are intrinsic properties of the system that is being modeled and not simply artifacts of the observer's own internal language. The other major problem is the appropriate interpretation of symbolic patterns learned from experience, which he refers to as proper interpretation of the current situation. Tani applied his model-based approach to the navigation of an experimental mobile robot. Tani's model, based on a recurrent neural network, demonstrated robot learning of symbolic structure in the geometry of the workspace. Furthermore, the robot generated diverse action plans to reach an arbitrary goal using the acquired model.

Binder and Pedraza[129] used the standard map model (a widely used model for Hamiltonian chaos) to produce the symbolic dynamics of a Poincaré section of a periodically kicked rotor. Their objective was to demonstrate that Markov tree and Markov chain models for such a physically realizable system can produce nonregular grammars near conditions where there is an order-chaos transition. (In this context, grammar refers to the Chomsky language hierarchy that is based on the amount of computational power needed to recognize that a particular string is an instance of the language in question). The 5-letter symbolization used was based on the motion of the map value on each iteration relative to special dynamical structures known as cantori. From their results, Binder and Pedraza conclude that both context-free and context-sensitive languages are produced by this system depending on initial conditions.

Baptista et al.[59] proposed a new communication technique based on deterministic dynamics modeling of language. They created a time-delay coarse-graining of the logistic map phase space based on the symbol-sequence statistics and then transmitted messages to a receiver by means of codewords that are specific phase-space targeting instructions rather than an explicit message. The authors reported that their approach yields error correction, compression, high security, and language recognition.

The combination of symbolization with neural network learning for noisy time series prediction is discussed by Giles et al.[130] The authors use a self-organizing map (SOM) to symbolize initial time-series data and then train recurrent neural networks to predict future symbols based on grammatical inference. The approach specifically takes advantage of the known abilities of such neural networks to learn deterministic grammatical rules that capture predictability in the evolution of the series. Giles et al. demonstrate their method using measured foreign exchange rate data.

G. Data Mining, Classification and Rule Discovery

The methods of data symbolization have also been applied for data mining, classification and rule discovery. Data mining is the process of finding useful information amidst a large set of information where patterns and rules might not be obvious. Classification is the process of building models from data for identification of unknown data sets. Rule discovery is the process of identifying rules of sequential patterns or relationships between patterns over time.

Typically, rule discovery relies on inherently symbolic data such as "people who buy eggs and root beer also buy avocados". In a seminal paper, Das et al.[131] applied rule-discovery techniques to real-valued time series via a process of symbolization. They symbolized the time-series data by clustering relative data relationships within a sliding time window and assigning symbols to each cluster (e.g., one symbol might represent three consecutively increasing data values, another might represent a low-high-low sequence of data, and so on). Thus, all relative data patterns were represented without defining partitions. (Note that this type of symbolization is similar to that employed by Bandt and Pompe[57] for their permutation entropy.) They then used the J-measure for rule ranking[132,133] and examined stock-price trends for significant rules.

André-Jönsson and Badal[134] used first differences to encode time-series patterns into a "Shape Description Alphabet" composed of letters. The magnitudes of differences between adjacent data records were significant in assigning symbols. The resulting encoded text stream was then indexed into a signature file and used for "blurry matching" of similar data patterns.

Self-organizing maps are often used for clustering continuous data, but for discrete data such as symbol strings, incremental learning laws are not easily defined. Kohonen and Somervuo[135] adapted the internal distance measures used in forming SOMs for use with symbol strings. They analyzed phoneme data of spoken Finnish using unsupervised learning as well as supervised improvements using Learning Vector Quantization.

Hebrail and Hugueney[136] also used clustering of small windows of time-series data to define non-partitioned symbolic descriptions of data patterns. They then used this symbolic description for visualization and for mining of sequential patterns. One key feature of the work was the compression of long time series into a few symbols representing distinctive episodes of time-series behavior

(for instance, a time series of 8760 records was compressed into a symbol-sequence representation of length 26).

An interesting clustering application is the phylogeny of historical texts, in which variant texts are examined for closeness to the original source based on a genetic-type analysis. Barbrook *et al.*[137] performed such an analysis on 44 15th-century texts of "The Wife of Bath's Prologue" from *The Canterbury Tales* by Geoffrey Chaucer. In this analysis, they treated distinct phrases ("characters") as genes, or symbols, and produced a phylogenetic tree using the split-decomposition software SplitsTree. They concluded that the original source was an annotated working draft with notes on additions or deletions.

Karp[138] adapted methods of pathway databases to encode scientic theories into symbolic form. Pathway databases are used to describe linked relationships such as the metabolic maps of bacteria. Karp argued that as scientific theories become more complex, the need for symbolically encoding the elements of the theory becomes more important.

VII. DISCUSSION

As interest continues to grow in high-speed data acquisition and observation of complex dynamical systems, symbolic analysis will clearly remain an important research tool. Symbolic methods offer a unique potential for computational efficiency, ease of visualization, and connections with information theory, language, and artificial intelligence that cannot be matched by any other approach.

Perhaps the most important (and contentious) current issue in application of these methods is the development of algorithms for appropriately defining symbols in the absence of generating partitions. In some cases new approaches for noise reduction may lead to more accurate empirical reconstruction of the underlying phase space. Clearly, however, there will always be many important cases in which experimentalists will need algorithms for defining symbols from noisy experimental measurements where little is known about the underlying physics. Even if the symbolic transformation always involves some degree of imprecision, the successful application of symbolization in numerous experimental contexts seems to indicate that such a goal is not unrealistic.

Closely associated with the question of symbol definition is the need to have efficient algorithms for defining appropriate symbol sequence (word) lengths. The recent development of approaches that automatically allow for variable sequence lengths (*e.g.*, context trees), and updating of these length with time, appears to offer considerable hope that such methods will be available soon.

As more engineering applications of symbolics emerge, we expect to see increased utilization of 'hard-wired' symbol transformation in which the conversion to symbols occurs directly in the measurement instrumentation. Thus post-processing and cost are minimized while speed is enhanced. Such applications are likely to be limited primarily to systems or processes in which the objective is to detect the onset of well understood dynamical transitions (e.g., bifurcation sequences) in real time.

VIII. ACKNOWLEDGEMENTS

[1] J. B. Green Jr., R. M. Wagner, and C. S. Daw, in *Proceedings, Spring Technical Meeting, Central States Section of The Combustion Institute* (Knoxville, Tennessee, 2002). Available from http://www-chaos.engr.utk.edu/pap/crg-cssci2002ice-paper.pdf.

[2] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (reprinted, University of Illinois Press,

Urbana and Chicago, Illinois, 1998).

[3] E. Seneta, *Non-Negative Matrices and Markov Chains* 2nd ed. (Springer-Verlag, New York, 1981).

[4] P. Baldi and S. Brunak, *Bioinformatics: The Machine Learning Approach* (MIT Press, Cambridge, Massachusetts, 1998).

[5] B. P. Kitchens, *Symbolic Dynamics: One-Sided, Two-Sided and Countable State Markov Shifts* (Springer, New York, 1998).

[6] J. Hadamard, J. Math. Pures Appl. **4**, 27 (1898).

[7] M. Morse, Trans. Amer. Math. Soc. **22**, 84 (1921).

[8] M. Morse and G. A. Hedlund, Amer. J. Math. **60**, 815 (1938).

[9] P. Collet and J. P. Eckmann, *Iterated Maps on the Interval as Dynamical Systems* (Birkhäuser, Basel, 1980).

[10] P. F. Holmes, Phys. Rep. **193**, 138 (1990).

[11] F. Diacu and P. Holmes, *Celestial Encounters: The Origins of Chaos and Stability* (Princeton University Press, Princeton, 1996).

[12] S. M. Ulam, *A Collection of Mathematical Problems* (Interscience Publishers, New York, 1960), see p. 73.

[13] A. Lasota and M. C. Mackey, *Chaos, Fractals and Noise: Stochastic Aspects of Dynamics* (Springer-Verlag, New York, 1991).

[14] A. B. Rechester and R. B. White, Phys. Lett. A **156**, 419 (1991).

[15] A. B. Rechester and R. B. White, Phys. Lett. A **158**, 51 (1991).

[16] G. Nicolis, in *Noise and Chaos in Nonlinear Dynamical Systems*, edited by F. Moss, L. A. Lugiato and W. Schleich (Cambridge University Press, New York, 1990), pp. 241–260.

[17] E. A. Jackson, *Perspective of Nonlinear Dynamics* (Cambridge University Press, New York, 1995).

[18] T. Bedford, M. Keane, and C. Series, *Ergodic Theory, Symbolic Dynamics and Hyperbolic Spaces* (Oxford Science Publications, Oxford, 1992).

[19] J. P. Crutchfield and N. H. Packard, Physica D **7**, 201 (1983).

[20] L. B. C. Cunningham and W. R. B. Hynd, J. Roy. Stat. Soc. Suppl. **8**, 96 (1946).

[21] F. C. Moon, *Chaotic and Fractal Dynamics* (John Wiley and Sons, New York, 1992).

[22] K. Steiglitz, *A Digital Signal Processing Primer* (Addison-Wesley, New York, 1996).

[23] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete Time Signal Processing* (Prentice Hall, New York, 1999).

[24] M. C. Cuéllar and P.-M. Binder, Phys. Rev E **64**, 046211 (2001).

[25] P. beim Graben, Phys. Rev. E **64**, 051104 (2001).

[26] H. D. I. Abarbanel, *Analysis of Observed Chaotic Data* (Springer, New York, 1996).

[27] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, 1997).

[28] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw, Phys. Rev. Lett. **45**, 712 (1980).

[29] F. Takens, Lect. Not. Math. **898**, 366 (1981).

[30] T. Sauer, J. Yorke, and M. Casdagli, J. Stat. Phys. **65**, 579 (1991).

[31] M. B. Kennel and A. I. Mees, Phys. Rev. E **61**, 2563 (2000).

[32] M. B. Kennel, Phys. Rev. E **56**, 316 (1996).

[33] T. Schreiber, Phys. Rev. Lett. **78**, 843 (1997).

[34] A. Witt, J. Kurths, and A. Pikovsky, Phys. Rev. E **58**, 1800 (1998).

[35] D. Yu, W. Lu, and R. G. Harrison, Chaos **9**, 865 (1999).

[36] T. Kapitaniak, K. Życzkowski, U. Feudel, and C. Grebogi, Chaos Sol. Fract. **11**, 1247 (2000).

[37] A. T. Hsu, A. G. Marshall, and T. L. Ricca, Anal. Chim. Acta **178**, 27 (1985).

[38] J. P. Crutchfield and K. Young, Phys. Rev. Lett. **63**, 105 (1989).

[39] P. Grassberger and H. Kantz, Phys. Lett. A **113**, 235 (1985).

[40] R. L. Davidchack, Y.-C. Lai, E. M. Bollt, and M. Dhamala, Phys. Rev. E **61**, 1353 (2000).

[41] M. Paluš, Biol. Cyber. **75**, 389 (1996).

[42] X. Z. Tang, E. R. Tracy, A. D. Boozer, A. deBrauw, and R. Brown, Phys. Rev. E **51**, 3871 (1995).

[43] X. Z. Tang, E. R. Tracy, and R. Brown, Physica D **102**, 253 (1997).

[44] P. E. Rapp, A. M. Albano, I. D. Zimmerman, and M. A. Jiménez-Montaño, Phys. Lett. A **192**, 27 (1994).

[45] L. M. Hively, P. C. Gailey, and V. A. Protopopescu, Phys. Lett. A **258**, 103 (1999).

[46] L. M. Hively, V. A. Protopopescu, and P. C. Gailey, Chaos **10**, 864 (2000).

[47] J.-S. Kim, J.-E. Park, J.-D. Seo, W.-R. Lee, H.-S. Kim, J.-I. Noh, N.-S. Kim, and M.-K. Yum, Phys. Med. Biol. **45**, 3403 (2000).

[48] J. Godelle and C. Letellier, Phys. Rev. E **62**, 7973 (2000).

[49] S. Kádár, J. Wang and K. Showalter, Nature **391**, 770 (1998).

[50] J. A. Freund, J. Kienert, L. Schimansky-Geier, B. Beisner, A. Neiman, D. F. Russell, T. Yakusheva, and F. Moss, Phys. Rev. E, **63**, 031910 (2001).

[51] H. A. Braun, M. Dewald, K. Schafer, K. Voigt, X. Pei, K. Dolan, and F. Moss, J. of Comp. Neuroscience, 7(1), 17 (1999).

[52] B. F. Feeny and F. C. Moon, Phys. Lett. A **141**, 397 (1989).

[53] E. M. Bollt, T. Stanford, Y.-C. Lai, and K. Życzkowski, Phys. Rev. Lett. **85**, 3524 (2000).

[54] E. M. Bollt, T. Stanford, Y.-C. Lai, K. Życzkowski, Physica D **154**, 259 (2001).

[55] M. Lehrman, A. B. Rechester, and R. B. White, Phys. Rev. Lett. **78**, 54 (1997).

[56] J. Kurths, U. Schwarz, A. Witt, R. Th. Krampe, and M. Abel, in *Chaotic, Fractal, and Nonlinear Signal Processing*, edited by R. A. Katz, Volume 375 of AIP Conference Proceedings (AIP Press, Woodbury, New York, 1996), pp. 33–54.

[57] C. Bandt and B. Pompe, Phys. Rev. Lett. **88**, 174102 (2002).

[58] R. Edwards, H. T. Siegelmann, K. Aziza, and L. Glass, Chaos **11**, 160 (2001).

[59] M. S. Baptista, E. Rosa Jr., and C. Grebogi, Phys. Rev. E **61**, 3590 (2000).

[60] J. S. Halow and C. S. Daw, AIChE Symp. Ser. **90**, 301 (1994).

[61] K. Mischaikow, M. Mrozek, J. Reiss, and A. Szymczak, Phys. Rev. Lett. **82**, 1144 (1999).

[62] S. Lesher, L. Guan, and A. H. Cohen, Neurocomputing **32–33**, 1073 (2000).

[63] C. S. Hsu, *Cell-to-Cell Mapping: A Method for Global Analysis of Non-Linear Systems* (Springer-Verlag, New York, 1987).

[64] B. Tombuyses and T. Aldemir, J. Sound Vib. **202**, 395 (1997).

[65] C. S. Daw, M. B. Kennel, C. E. A. Finney, and F. T. Connolly, Phys. Rev. E **57**, 2811 (1998).

[66] R. L. Devaney, *An Introduction to Chaotic Dynamical Systems, 2nd Edition* (Addison-Wesley, New York, 1989).

[67] M. B. Kennel and A. I. Mees, in *Nonlinear Dynamics and Statistics*, edited by A. I. Mees (Birkhäuser Verlag, Boston, 2001), pp. 387–412.

[68] T. Schürmann and P. Grassberger, Chaos **6**, 414 (1996).

[69] G. D'Alessandro and A. Politi, Phys. Rev. Lett. **64**, 1609 (1990).

[70] A. M. Fraser and H. L. Swinney, Phys. Rev. A **33**, 1134 (1986).

[71] W. Li, J. Stat. Phys. **60**, 823 (1990).

[72] R. F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).

[73] H. Herzel and I. Große, Physica A **216**, 518 (1995).

[74] R. M. Goldstein, IRE Transactions on Space Electronics and Telemetry, **170** (1962).

[75] M. S. Roulston, Physica D **125**, 285 (1999).

[76] B. Kedem, *Time Series Analysis by Higher Order Crossings* (IEEE Press, New York, 1994).

[77] J. H. van Vleck, RRL Report **51**, July 21 (1943).

[78] J. L. Lawson and G. E. Uhlenbeck, *Threshold Signals* (McGraw-Hill, New York, 1950).

[79] S. O. Rice, Bell System Tech. J. **23**, 282–332 (1944) and **24**, 46–156 (1945), reprinted in *Selected papers on noise and stochastic processes*, N. Wax, ed. (Dover, New York, 1954).

[80] J. Barnett and B. Kedem, IEEE Trans. Inf. Theory **44**, 1672-1677 (1998).

[81] X. Z. Tang and E. R. Tracy, Chaos **8**, 688 (1998).

[82] G. A. Darbellay in *Signal Analysis and Prediction*, edited by A. Procházka, J. Uhlíř, P. J. W. Rayner, and N. G. Kingsbury (Birkhäuser, Boston, 1998), pp. 249–262.

[83] F. C. Moon, W. Holmes, and P. Khoury, Chaos **1**, 65 (1991).

[84] C. S. Daw, J. B. Green Jr., R. M. Wagner, C. E. A. Finney, F. T. Connolly, *Proceedings of The Combustion Institute* **28**, 1249 (2000).

[85] T. Schreiber, Phys. Rev. Lett. **85**, 461 (2000).

[86] S. Pincus, Chaos **5**, 110 (1995).

[87] N. Perry and P.-M. Binder, Phys. Rev. E **60**, 459 (1999).

[88] P. E. Rapp, C. J. Cellucci, K. E. Korslund, T. A. A. Watanabe, and M. A. Jiménez-Montaño, Phys. Rev. E **64**, 016209 (2001).

[89] H. Herzel, Syst. Anal. Model. Simul. **5**, 435 (1988).

[90] H. Herzel, A. O. Schmitt, and W. Ebeling, Chaos Sol. Fract. **4**, 97 (1994).

[91] P. Grassberger, Phys. Lett. A **128**, 369 (1988).

[92] P. Grassberger, Physica A **140**, 319 (1986).

[93] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, Physica D **58**, 77 (1992).

[94] J. Theiler and D. Prichard, Physica D **94**, 221 (1996).

[95] T. Schreiber and A. Schmitz, Phys. Rev. Lett. **77**, 635 (1996).

[96] K. Dolan, A. Witt, M. L. Spano, A. Neiman, and F. Moss, Phys. Rev. E **59**, 5235 (1999).

[97] D. N. Chorafas, *Statistical Processes and Reliability Engineering* (Van Nostrand, Princeton, 1960).

[98] U. Schwarz, A.O. Benz, J. Kurths, and A. Witt, Astron. Astrophysics **277**, 215 (1995).

[99] M. J. van der Heyden, C. G. C. Diks, B. P. T. Hoekstra, and J. DeGoede, Physica D **117**, 299 (1998).

[100] G. Weiss, J. Appl. Prob. **12**, 831 (1975).

[101] C. Diks, J. C. van Houwelingen, F. Takens, and J. De-Goede, Phys. Lett. A **201**, 221 (1995).

[102] L. Stone, G. Landan, and R. M. May, Proc. R. Soc. Lond. B **263**, 1509 (1996).

[103] H. Voss and J. Kurths, Phys. Rev. E **58**, 1155 (1998).

[104] C. S. Daw, C. E. A. Finney, and M. B. Kennel, Phys. Rev. E **62**, 1912 (2000).

[105] V. V. Gavrishchaka and S. B. Ganuli, J. Geophys. Res. **106**, 6247 (2001).

[106] P. E. Greenwood, L. M. Ward, D. F. Russell, A. Neiman, and F. Moss, Phys. Rev. Lett. **84**, 4773 (2000).

[107] R. Steuer, W. Ebeling, D. F. Russell, S. Bahar, A. Neiman, and F. Moss, Phys. Rev. E **64**, 061911 (2001).

[108] F. Wendling, J.-J. Bellanger, J.-M. Badier, and J.-L. Coatrieux, IEEE Trans. Biomed. Engr. **43**, 990 (1996).

[109] P. beim Graben, J. D. Saddy, M. Schlesewsky, adn J. Kurths, Phys. Rev. E **62**, 5518 (2000).

[110] P. I. Saparin, W. Gowin, J. Kurths, and D. Felsenberg, Phys. Rev. E **58**, 6449 (1998).

[111] J. Kurths, A. Voss, P. Saparin, A. Witt, H. J. Kleiner, and N. Wessel, Chaos **5**, 88 (1995).

[112] A. R. Rao and R. C. Jain, IEEE Trans. Pattern Analysis & Machine Intelligence **14**, 693 (1992).

[113] E. N. Lorenz, in *Predictability, Proceedings of a Seminar at the European Center for Medium Range Weather Forecasts* (Reading, England, 1995), vol. I, pp. 1–18.

[114] E. Mazzucato and R. Nazikian, Phys. Rev. Lett. **71**, 1840 (1993).

[115] A. B. Rechester, M. Lehrman, R. S. Granetz, P. Stek, T. Evans, R. B. White, and E. Mazzucato, Bull. Am. Phys. Soc. **41**, 1457 (1996).

[116] M. Lehrman and A. B. Rechester, Phys. Rev. Lett. **87**, 164501 (2001).

[117] P. Angeli and G. F. Hewitt, in *Proceedings of the ASME Heat Transfer Division*, Volume HTD 334-3, American Society of Mechanical Engineers (1996), pp. 149–156.

[118] R. van der Welle, Int. J. Multiphase Flow **11**, 317, 1985.

[119] W. M. Gonçalves, R. D. Pinto, and J. C. Sartorelli, Physica D **134**, 267 (1999).

[120] P. Jung, J. Wang, R. Wackerbauer, and K. Showalter, Phys. Rev. E **61**, 2095 (2000).

[121] F. C. Moon, Chaos **1**, 31 (1991).

[122] R. Kobes, J. Liu, and S. Peleš, Phys. Rev. E **63**, 036219 (2001).

[123] R. E. Kalman, in *Proceedings of the Symposium on Nonlinear Circuit Analysis, Vol. 6* (Polytechnic Institute of Brooklyn, New York, 1956).

[124] D. F. Delchamps, IEEE Trans. Auto. Control AC-35, 916 (1990).

[125] E. M. Bollt and M. Dolnik, Phys. Rev. E **55**, 6404 (1997).

[126] M. Dolnik and E. M. Bollt, Chaos **8**, 702 (1998).

[127] S. D. Goodman and M. A. Brooke, Appl. Opt. **32**, 752 (1993).

[128] J. Tani, IEEE Trans. Systems, Man, & Cybernetics, Part B: Cybernetics **26**, 412 (1996).

[129] P.-M. Binder and J. M. Pedraza, Phys. Rev. E **62**, R5883 (2000).

[130] C. L. Giles, S. Lawrence, and A. C. Tsoi, in *Proceedings of IEEE/IAFE Conference on Computational Intelligence for Financial Engineering (CIFEr)*, IEEE, Piscataway, New Jersey, pp. 253–259 (1997).

[131] G. Das, K.-I. Lin, H. Mannila, G. Renganathan, and P. Smyth, *Proc. 4th Int. Conf. Rule Discovery and Data Min-*

ing (KDD-98), pp. 16–22 (1998).

132 P. Smyth and R. M. Goodman, in *Knowledge Discovery in Databases*, MIT Press, pp. 159–176 (1991).

133 P. Smyth and R. M. Goodman, IEEE Trans. Knowledge and Data Eng. **4**, 301 (1992).

134 H. André-Jönsson and D. Z. Badal, *Proc. Principles of Data Mining and Knowledge Discovery (PKDD-97)*, published as *Lecture Notes in Artificial Intelligence* **1263**, Springer, pp. 211–220 (1997).

135 T. Kohonen and P. Somervuo, Neurocomputing **21**, 19

136 (1998).

136 G. Hebrail and B. Hugueney, in Workshop on Symbolic Data Analysis, Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-2000), Lyon, France, 2000. Available from http://www.univ-lyon2.fr/~pkdd2000/Download/WS6_4.pdf.

137 A. C. Barbrook, C. J. Howe, N. Blake, and P. Robinson, Nature **394**, 839 (1998).

138 P. D. Karp, Science **293**, 2040 (2001).