SYMBOLIC DYNAMICS FROM BIOLOGICAL TIME SERIES

A. DESTEXHE

Service de Chimie Physique, Université Libre de Bruxelles, CP231 Campus Plaine, Blvd. Triomphe, B-1050 Brussels, Belgium

Received 2 June 1989; revised manuscript received 22 November 1989; accepted for publication 22 November 1989
Communicated by A.P. Fordy

A method is outlined for estimating the order of a Markov process, and applied to both model and experimental time series. Unexpected high values are seen for heart and brain recordings, showing that for these systems the succession of time intervals between patterns exhibits intimate correlations.

Chaotic dynamical systems share many common properties with stochastic processes. For example, the divergence of nearby trajectories imposes a limit on the predictability of the time behavior of the system beyond a characteristic time which is of the order of the inverse of the Kolmogorov entropy \cite{1,2}. The deterministic description of the attractor as a set of trajectories in the phase space can then be replaced by a statistical description in terms of a probability density in the phase space \cite{2}.

On the other hand, chaotic dynamics contains remarkable regularities that are not usually seen in noise-driven systems. The same regions of the phase space may be visited at various times during the dynamical evolution of the system. In some cases, the system’s dynamics may consist of an irregular succession of distinguishable events corresponding to a visit of the trajectory in a well defined region of the phase space. For example in the Lorenz model \cite{3}, the chaotic attractor is composed of two identical regions, symmetrically disposed around the z-axis. Despite the deterministic character of the Lorenz equations, these regions are visited by the trajectories in a stochastic manner \cite{4}.

In order to describe the succession of events in such systems, it is useful to associate a symbol to each one of these events or, in other words, to partition the phase space into a finite set of regions each associated with a different symbol. This operation transforms the dynamics of continuous variables into a discrete sequence of symbols whose statistical properties and time correlations can be studied. In particular, it is of interest to consider such a sequence of symbols as a Markov chain. The order of the corresponding Markov process will give useful information about the time correlations of the underlying dynamical system.

In this paper, we will show how to use the symbolic description to assess information about the dynamics of an experimental system, usually known through a single time series representing the evolution of one of its variables. First, we will describe a method, based on standard statistical concepts, to evaluate the order of a Markov chain with a good accuracy. This method will be illustrated on two model systems, then it will be used to analyze the sequence of symbols extracted from two biological signals recorded from the human heart and the human brain. In these systems, the dynamics is made of the succession of a given pattern of activity and the time intervals between these patterns will be “translated” into symbols proportional to their duration. The sequence obtained will reflect the succession of intervals lengths and the order of the Markov chain will provide an interesting insight into the correlations among successive patterns.

For the Lorenz attractor, Aizawa \cite{4} associated the letter L (respectively R) to each orbit if the trajectory was on the left (respectively right) side of the attractor. The sequence of symbols reflects the succession of visits of the trajectory in either side of the attractor. He showed that for a given set of pa-
rameters the sequence of symbols may be a zeroth order Markov process (Bernoulli process [5]). In other words, a deterministic system described by three coupled nonlinear differential equations of the first order may generate a sequence of totally uncorrelated symbols, as if they were generated by a stochastic process such as the coin tossing. Nicolis et al. [6,7] considered the three-variable Rössler model in the chaotic region. Each time a variable crosses a prescribed threshold, a symbol specific to that variable is produced. The sequence of these symbols will then provide the image of the various thresholds encountered successively by the dynamics. They showed that the associated Markov process is at least of the fifth order. At the opposite of the Lorenz model, the dynamics of threshold crossing in the Rössler model may generate highly correlated sequences of symbols.

Following the ideas of Billingsley [8], Hoel [9] and Nicolis et al. [6,7], we now outline a method for estimating the order of a Markov process. Let $X_1, X_2, ..., X_N$ be the sequence of $N$ symbols describing the system’s dynamics. The symbols are chosen among the $n$ different letters of an alphabet $A$ ($n$ is also the number of states of the process). The probability of observing in the sequence the word $X_iX_{i+1}X_{i+2}...X_k$ is $P_i = P(x_k | x_{k-1} | ... | x_{i+1})$. Here the subscript $i$ indicates that $X_iX_{i+1}X_{i+2}...X_k$ is the $i$th word among the $n^k$ possible words of length $k$. The words are sorted in alphabetical order. For a Markov process of the $k$th order, the conditional probability of observing $X_k$ as the $k$th symbol in the word obeys the relation [10]

$$P(X_k | X_1X_2X_3...X_{k-1}) = P(X_k | X_{k-1}) ,$$

(1)

The above two equations can be calculated from the sequence and compared for different values of $k$ until the values coincide for all possible words. It is also useful to compare $P(X_1X_2X_3...X_k)$ with $P^{(k)}(X_1X_2X_3...X_k)$ which is the probability of observing the sequence $X_1X_2X_3...X_k$ deduced by assuming a $k$th order Markov chain. $P^{(k)}(X_1X_2X_3...X_k)$ is calculated from the conditional probabilities as follows. From the definition of conditional probability [5], we can write for different values of $k$

$$P^{(k)}(X_1X_2X_3...X_k) = P(X_k)P(X_{k-1})...P(X_1) ,$$

(2)

$$P^{(1)}(X_1X_2X_3...X_k) = P(X_k)P(X_{k-1})...P(X_1) ,$$

(3)

$$P^{(2)}(X_1X_2X_3...X_k) = P(X_k)P(X_{k-1})...P(X_1) ,$$

(4)

etc.

The values $P(X_1X_2X_3...X_k)$ deduced from counting are compared with the computed values of $P^{(k)}(X_1X_2X_3...X_k)$ by using a test statistic such as

$$\chi^2 = \frac{n}{n^k} \sum_{i=1}^{n^k} \left( P_i - P^{(k)}_i \right)^2 / P_i .$$

(5)

The procedure is repeated for increasing values of $k$ and $\chi^2$ must converge to zero for $k > k_0$, where $k_0$ is the order of the Markov process.

However in experimental situations when the amount of symbols in the data set is limited, the convergence of $\chi^2$ to zero may be ambiguous. We propose here a method to clear this ambiguity. The quantity evaluated numerically from counting is the frequency $f'$ which is defined as

$$f'(x_1x_2x_3...x_k) = \frac{\text{[number of words } x_1x_2x_3...x_k]}{\text{[number of words of length } n^k]} ,$$

(6)

and is equal to the probability $P(x_1x_2x_3...x_k)$ when $N \to \infty$. If we assume that the frequencies are converging to the probabilities as a negative power of the length $N$, such as

$$f(N) = P + b_N N^{-\alpha} , \quad \alpha > 0 ,$$

(7)

then replacing in (5) the probabilities $P_i$ by the frequencies as defined in (7) leads for large $N$ and for $k > k_0$ ($P_i = P^{(k)}_i$) to

$$\chi^2 = (N^{-2\alpha} / n^k) \sum_{i=1}^{n^k} (b_i - b_i^{(k)})^2 / P_i .$$

(8)

We see that for large values of $N$, $\chi^2$ may have a qualitatively different behavior depending on whether $k < k_0$ or $k > k_0$: in the first case, $\chi^2$ will be independent of $N$ and in the second case, $\chi^2$ is proportional to a negative power of $N$. This can be generalized to other forms of dependence on $N$. If more generally, the frequency can be expressed as the probability plus
a function of $N$, the same function of $N$ is seen in $\chi^2$ for $k \geq k_0$ while for $k < k_0$, $\chi^2$ remains independent of $N$. This property can be used to discriminate between the two cases, providing a criterion to estimate $k_0$. In other words, we propose to use the abrupt change of the scaling behavior of $\chi^2$ with $n$ to estimate the order of the process.

The abrupt change of scaling of $\chi^2$ observed for large $N$ is illustrated in fig. 1a where $\chi^2$ is represented as a function of $N$ for a third order Markov process. We see that $\chi^2$ scales clearly as a negative power of $N$ for $k \geq k_0$ ($\alpha = 0.51 \pm 0.05$ for $k = 3$ to 5), therefore our conjecture (7) seems to hold in this case. A similar scaling was found for the tent map and the logistic map ($r=4$). The order estimated was respectively of zero and one using equipartitions in two and four symbols for both maps. It is to be noticed that a similar scaling behavior with a slope close to 0.5 has been observed previously [11] for the excess entropy which scales like $M^{-n}$ where $M$ is the number of possible sequences ($M=n^k$).

In the particular case of the Rössler model (as described in refs. [6,7]), we evaluated the slope of log $\chi$ versus log $N$ (fig. 1b) and we have found a change of scaling behavior at a value of $k_0=5$ which confirms previous estimates [6,7]. Other forms of statistic test can be used, such as for example the distance between the distributions $\{P_i\}$ and $\{P_i^{(n)}\}$ in the $n$-dimensional space. We verified that the same behavior was found with the Euclidian distance and the Hamming distance.

The symbolic dynamics may also be used to analyze systems described by continuous variables when a single recognizable event appears at irregular time intervals. The aim of the following paragraphs is to show how to transform the dynamics of two biological processes into a sequence of symbols and then use the method described above to assess the order of the corresponding Markov chain. The order of the Markov process may bring information on how far the successive occurrences of the event are temporally related, a value of zero indicating a complete statistical independence between these occurrences.

Fig. 2 depicts the time evolution of two biological signals for which a characteristic event is produced irregularly. The first signal (fig. 2a) represents the electrocardiogram (ECG) of a normal human heart and is measured at the level of the thorax. The second signal (fig. 2b) shows the electrical activity of the human brain (electroencephalogram or EEG) recorded during a state of deep coma (a degenerative disease of the brain known as the Creutzfeldt–Jakob disease). The recording of the electrical potentials and the digitization were performed using standard techniques which will not be described here (more details can be found in ref. [12] for fig. 2a and in ref. [13] for fig. 2b). The reconstruction of phase portraits and the evaluation of the correlation dimension [12,13] or Lyapunov exponents [12,14]

![Diagram](image-url)
Fig. 2. (a) Electrical activity of the heart (ECG) recorded at the level of the thorax from a normal human subject. (b) Electrical activity of the human brain (EEG) during a pathological state. Both signals exhibit a repetition of the same pattern at irregular intervals.

all point to the same conclusion, namely that the electrical activity of the heart as well as the brain waves may follow chaotic dynamics.

From fig. 2, it is seen that both for ECG and EEG recordings, electric potential spikes are produced at irregular time intervals. In the case of the EEG, the distribution of inter-spike time intervals may be bimodal. Associating a symbol to each peak of the distribution is equivalent to studying the succession of “short” and “long” intervals. Therefore, in this case, it is natural to associate a symbol to the duration of the interval. At the opposite of a partition determined from the distribution of intervals, we may consider an equipartition of the interval. The order of the process has been found to be insensitive to either of these partitions. In the following paragraphs, the intervals will be equipartitioned in a few subsets depending on the duration and each subset will be associated with a prescribed symbol.

Fig. 3 displays the values of $\chi^2$ obtained respectively from the sequences of ECG (fig. 3a) and EEG intervals (fig. 3b). The order of the corresponding Markov process deduced from the convergence of $\chi^2$.

Fig. 3. Test statistic $\chi^2$ versus order $k$ for the sequences of symbols extracted from the signals of fig. 2: (a) ECG ($N=2132$), (b) EEG ($N=1756$). Several different partitions of the interval in $n$ symbols have been considered: $n=2$ (+), $n=3$ (x), $n=4$ (triangles) and $n=5$ (squares). (c) Third order Markov process for comparison (same parameters as fig. 1).
is of $k_0 = 3$ for ECG intervals (see fig. 3c for comparison). In the case of the EEG intervals, $\chi^2$ converges to zero at $k_0 = 6$ (fig. 3b). The convergence of $\chi^2$ to zero does not depend significantly on the number $n$ of symbols used. Actually, different values of $n$ were tested (from $n = 2$ to $n = 5$) and give similar results. It is remarkable that the order of the Markov process obtained from the symbolic dynamics generated with "short" and "long" intervals does not differ significantly from the dynamics presented with the attractor and the order of the time interval Markov chain has not yet been established. A high order Markov process may be reminiscent of chaotic dynamics, such as for example in the Rössler model. More generally it has been shown [16] that a given class of chaotic dynamical systems can be mapped into a well defined stochastic process described by a master equation. In some cases, this process can be of higher order.

The fact that the crossing beat-to-beat intervals may
ECG (3 pseudo-cycles), the autocorrelation function vanishes more rapidly for the EEG ($t_c = 2.6$ s or 3.8 pseudo-cycles) than for the ECG ($t_c = 5$ s or 4.4 pseudo-cycles). This result shows that symbolic dynamics may bring complementary information to usual techniques.

The symbolic analysis performed here is independent of any phase space reconstruction algorithm, and does not suffer from the limitations inherent to the reconstruction of high dimensional attractors [22,23]. Moreover, the intervals are subject to a coarse “digitization” into a few symbols. Therefore, we expect that the influence of additive noise will be insignificant (the amplitude of the noise is usually much smaller than the typical order magnitude of the partition of intervals).

In conclusion, in this paper we have outlined a method, based on standard statistical concepts, to evaluate the order of a Markov chain obtained from a time series. The scaling properties of the $\chi^2$ test also for stimulating discussions. This work has been achieved under I.R.S.I.A. fellowship.

References