Testing independence in time series via universal distributions of permutations and words

Camillo Cammarota [*]   Enrico Rogora [†]

Abstract

We study probability distributions of permutations and binary words, which arise in symbolic analysis of time series and their differences. Under the assumptions that the series is stationary and independent we show that these probability distributions are universal and we derive a recursive algorithm for computing the distribution of binary words. This provides a general framework for performing chi square tests of goodness of fit of empirical distributions versus universal ones. We apply these methods to analyze heartbeat time series; in particular we measure the extent to which atrial fibrillation can be modeled as an independent sequence.

Keywords: Symbolic dynamics, time series, binary words, permutations, ECG analysis.
Mathematics Subject Classification 2000: 62P10, 92C55

1   Introduction

In time series analysis it can be very useful to consider a finite partition of the state space on which to project the dynamics. If the elements of the partition are identified with symbols from a finite alphabet (e.g letters or digits), the projected time series is coded into a sequence of symbols and the projected dynamics is called *symbolic dynamics*. Much attention has been devoted to the case where the alphabet consists only of two symbols, say 0 and 1. In this case, any segment of a time series is coded into a binary word of the same length. An important object of study is the probability distribution induced by the coding on the space of binary

[*]Dept. of Mathematics, University of Rome "La Sapienza", cammar@mat.uniroma1.it
[†]Dept. of Mathematics, University of Rome "La Sapienza", rogora@mat.uniroma1.it

words of given length. It is also possible to code the order relationships between values in a segment of the time series into the permutation which reorders the segment. This gives a finer encoding which, for some purposes, is more enlightening than the binary one. Here also an important object of study is the probability distribution induced by the permutation coding on the space of permutations and the relationships with the probability distribution on binary words.

Symbolic dynamics on binary words and permutations can be useful in order to distinguish the behavior of time series. In particular this can be done for time series which arise from data measurements of complex systems for which there is often not a satisfactory model. This happens generally in physiological data and in particular in the heartbeat time series, to which our analysis is also concerned. In literature greater attention has been devoted to the coding of these series in binary words in various physiological and pathological conditions. Various approaches have been followed.

A first approach is based on coding the difference between two consecutive values of the series since the differentiated series is known to be more stationary than the original one (see for instance [1]). Depending on the type of application, it can be suitable to fix a threshold under which the difference is not considered significant. For this type of coding, we refer to [5] and [11]. Otherwise no threshold is considered as in [2], [3], [6], [7]. A second approach is based on coding directly the original signal as in [12] and [13]. Also permutation patterns have received some attentions as a method of coding physiological signal, as in [8] and [10].

In the analysis of signals, when a deterministic model is not available, a probabilistic approach is often used and the time series is modeled as a (usually stationary) sequence of random variables. In order to characterize the dependence between the elements of the series it is important to quantify how the series differs from a sequence of independent variables. One of the methods which has been used is to shuffle the series in order to get a new series which has the same one dimensional distribution but no correlation; the measure of the distance from independence is performed by comparing the occurrence frequency of binary words in the original time series and in the shuffled one [6].

In this paper we introduce a general framework under which some of the above methods can be subsumed. Our approach is based on the properties of permutations and words distributions under the assumption of independence. We consider a stationary sequence of independent random variables (i.i.d.r.v.) and the sequence of their differences, which is still stationary but not independent. We also consider the distributions of permutations and words associated respectively to the sequence and to the differences. These distributions are universal, i.e. they do not depend on the distribution of the variables. It is well known that the distribution of permutations is uniform but, at our knowledge, the distribution of words

has never been investigated, and we provide an explicit algorithm for computing it.

On the side of data analysis, we consider the empirical distributions of permutations and words occurring in our time series. In order to quantify the distance of the data series from an i.i.d.r.v. sequence we compare the empirical distribution with the universal ones. This can be done by using the standard chi square test of goodness of fit.

As an example we apply our method to the analysis of the 24 hours heartbeat time series. We consider two cases of atrial fibrillation and two normal subjects taken from the database of our research group, already used in the papers quoted above. Although the four series are all significantly non independent, the chi square values of fibrillating are much closer to independent i.i.d. series than to normals (see table 2). This result seems to agree with the one in [4] obtained with a completely different method, and with that in [6] obtained with the shuffling method. In particular, in our previous paper [4] the analysis of time series was performed by taking explicitly into account the non stationary behavior of the heart rate with the estimation of the trend and the analysis of the residuals. This was necessary since the technique used there to investigate the dependence heavily relied on stationarity assumptions. On the contrary, we believe that the technique here introduced, since distribution free, is less dependent on these assumptions.

2 Probability distributions on permutations and binary words

For coding a segment of a time series into a permutation we use the following method. In $\mathbb{R}^{n+1}$ let us consider the subset Δ (called *big diagonal*) which is the set of points $(x_1, \ldots, x_{n+1})$ for which there exist at least two indexes i, j such that $x_i = x_j$ and denote $\mathbb{R}^{n+1} \setminus \Delta$ by $\mathbb{R}^{n+1}_{\neq}$. Let S_{n+1} denote the symmetric group, i.e. the set of permutations of $\{1, \ldots, n+1\}$. We use one line notation for permutations, i.e.

$$(i_1, i_2, \ldots, i_{n+1})$$

denotes the permutation

$$\begin{pmatrix} 1 & 2 & \ldots & n+1 \\ i_1 & i_2 & \ldots & i_{n+1} \end{pmatrix}$$

We define a function $\Pi : \mathbb{R}^{n+1}_{\neq} \to S_{n+1}$ in the following way:

$$\Pi(x_1, \ldots, x_{n+1}) = (\pi(1), \ldots, \pi(n+1))$$

3

where

$$\pi(i) = 1 + \#\{j : x_j < x_i\}$$

(The symbol # denotes cardinality). Note that with this definition, if π^{-1} is the inverse of π, then

$$x_{\pi^{-1}(1)} < x_{\pi^{-1}(2)} < \cdots < x_{\pi^{-1}(n+1)}$$

For example, if $(x_1, \ldots, x_4) = (181, 32, 42, 115)$, then $\pi = (4, 1, 2, 3)$, $\pi^{-1} = (2, 3, 4, 1)$ and

$$x_2 < x_3 < x_4 < x_1.$$

For coding a segment of a time series into a binary word we use the following method.

If $\mathbb{Z}_2$ denotes the set $\{0, 1\}$, we call $\mathbb{Z}_2^n$ the set of binary words of length n.

We define a function $W : \mathbb{R}_{\neq}^{n+1} \to \mathbb{Z}_2^n$ in the following way:

$$W(x_1, \ldots, x_{n+1}) = (w(1) \ldots, w(n))$$

where

$$w(i) = \begin{cases} 0, & \text{if } x_i > x_{i+1} \\ 1, & \text{if } x_i < x_{i+1} \end{cases} \tag{1}$$

For example

$$W(181, 32, 42, 115) = (0, 1, 1)$$

Note that for each $x = (x_1, \ldots, x_{n+1})$,

$$W(\Pi(x)) = W(x) \tag{2}$$

This is the reason why we have chosen π and not its inverse in the definition of Π. The problem to compute how many permutations correspond to the same word will be an object of our investigation.

Let us denote by $X = (X_1, \ldots, X_{n+1})$ a collection of $n + 1$ independent random variables identically distributed with probability density f positive and absolutely continuous with respect to the Lebesgue measure of $\mathbb{R}$, and let P be the product probability measure on $\mathbb{R}^{n+1}$. We consider the subsets

$$A_\pi = \{(x_1, \ldots, x_{n+1}) \in \mathbb{R}_{\neq}^{n+1} : \Pi(x_1, \ldots, x_{n+1}) = \pi\}$$

parametrized by the permutations $\pi \in S_{n+1}$. Note that A_π can be identified with the event

$$\{X_{\pi^{-1}(1)} < X_{\pi^{-1}(2)} < \cdots < X_{\pi^{-1}(n+1)}\}$$

4

The subsets A_π give a partition of $\mathbb{R}_{\neq}^{n+1}$. We define the probability measure P_Π over S_{n+1} induced by P as

$$P_\Pi(\pi) = P(A_\pi)$$

Lemma (see [9]) The probability P_Π is uniform on S_{n+1}, i.e.

$$P_\Pi(\pi) = 1/(n+1)!$$

We now consider the probability measure induced by P on $\mathbb{Z}_2^n$ or equivalently by P_Π in force of (2). For any word w we define the subset

$$B_w = \{(x_1, \dots, x_{n+1}) \in \mathbb{R}_{\neq}^{n+1} : W(x_1, \dots, x_{n+1}) = w\}$$

and we define the induced probability P_W by

$$P_W(w) = P(B_w)$$

We consider an explicit example.

Permutations		Words
$(1,2,3)$	$\rightarrow$	11
$(1,3,2)$	$\rightarrow$	10
$(2,1,3)$	$\rightarrow$	01
$(2,3,1)$	$\rightarrow$	10
$(3,1,2)$	$\rightarrow$	01
$(3,2,1)$	$\rightarrow$	00

The probability induced over words of length n can be simply computed by counting all permutations which are mapped to a given word and dividing by $1/(n+1)!$. In the example we have

$$P_W(11) = 1/6, \ P_W(10) = P_W(01) = 1/3, \ P_W(00) = 1/6$$

In general, the same kind of argument proves that P_W is explicitly computable by P_Π and does not depend on the probability density f.

3 Computation of P_W

We now show how to compute directly P_W in the i.i.d. case without referring to P_Π and this can be useful when we analyze long words. This computation is a recursive one based on the following framework.

We first introduce some notations. By $|w|$ we denote the length of a binary word, by 1^n we denote the word of length n containing only 1's, by vw we denote the

word obtained by concatenating v and w, by $v \bullet w$ we denote the two element set $\{v0w, v1w\}$ and by $()$ we denote the empty word, i.e the only word in $\mathbb{Z}_2^0$.
Let $\{\mathcal{P}_n\}_{n \geq 0}$ be a family of probability distributions, where $\mathcal{P}_n$ is defined over $\mathbb{Z}_2^n$ such that the following properties hold

1. $\mathcal{P}_0(()) = 1$

2. $\mathcal{P}_{|v|+1+|w|}(v \bullet w) = \mathcal{P}_{|v|}(v)\mathcal{P}_{|w|}(w)$

When it is clear from the context we shall drop the subindex from $\mathcal{P}_i$.
Note that, since the $\mathcal{P}$'s are probabilities, one always has

$$(*) \qquad \mathcal{P}(w0) + \mathcal{P}(w1) = \mathcal{P}(w\bullet).$$

Theorem If the family of probabilities $\{\mathcal{P}_n\}$ satisfies properties 1 and 2 and $\mathcal{P}_n(1^n)$ are known for all n's, then they are completely determined.

Proof The proof consists of an explicit algorithm to compute $\mathcal{P}(w)$.
For the empty word one has $\mathcal{P}(()) = 1$ by property 1.
For words of length 1, $\mathcal{P}(1)$ is assumed to be known and $\mathcal{P}(0) = 1 - \mathcal{P}(1)$.
In general, given a binary word w, let us assume that we already know how to compute $\mathcal{P}(u)$ for all words of length less than $|w|$. If $w = 1^n$ then $\mathcal{P}(1^n)$ is known. Otherwise let $w = u0v$ where v contains no 0's. Then $\mathcal{P}(w) = \mathcal{P}(u \bullet v) - \mathcal{P}(u1v)$ by (*). Moreover $\mathcal{P}(u \bullet v) = \mathcal{P}(u)\mathcal{P}(v)$ by property 2 and we are able to compute the two factors $\mathcal{P}(u)$ and $\mathcal{P}(v)$ since they have length shorter than that of w. To compute $\mathcal{P}(u1v)$ we consider the last zero in u and we decompose the word as before. By iterating the procedure we finally get words with only 1's. $\qquad \square$

An example will make the proof of the theorem clearer.

$$\mathcal{P}(1010) = \mathcal{P}(101\bullet) - \mathcal{P}(1011) = \mathcal{P}(101) - \mathcal{P}(1011) =$$
$$\mathcal{P}(1 \bullet 1) - \mathcal{P}(111) - \mathcal{P}(1 \bullet 11) + \mathcal{P}(1111) =$$
$$\mathcal{P}(1)\mathcal{P}(1) - \mathcal{P}(111) - \mathcal{P}(1)\mathcal{P}(11) + \mathcal{P}(1111) =$$
$$\mathcal{P}(1)^2 - \mathcal{P}(1^3) - \mathcal{P}(1)\mathcal{P}(1^2) + \mathcal{P}(1^4)$$

Corollary The algorithm given in the proof of the theorem can be used to compute P_W.

Proof The i-th symbol $w(i)$ in the word w is a random variable which depends only on the difference $X_{i+1} - X_i$. Hence $w(i)$ and $w(i+2)$ are independent in force of the independence of the variables X_i. This implies property 2. Note however that P_W is not trivial since $w(i)$ and $w(i+1)$ are not independent.

	000	001	010	011	100	101	110	111
00	0.001	0.007	0.019	0.014	0.026	0.049	0.037	0.014
01	0.019	0.056	0.085	0.049	0.036	0.056	0.026	0.007
10	0.007	0.026	0.056	0.036	0.049	0.085	0.056	0.019
11	0.014	0.036	0.049	0.026	0.014	0.019	0.007	0.001

Table 1: Rounded values of probability distribution on length 5 binary words. The first two digit of the word are those in the first column and the last three are those in the first row. e.g. the word 01111 has probability 0.007.

Since P_W is induced by P_Π

$$P_W(1^n) = P_\Pi(1, 2, \ldots, n+1) = \frac{1}{(n+1)!}$$

For $n = 0$ this gives property 1. $\square$

As an example we compute $P_W(1010)$, completing the computations of the previous example.

$$P_W(1010) = P_W(1)^2 - P_W(1^3) - P_W(1)P_W(1^2) + P_W(1^4) =$$

$$\left(\frac{1}{2}\right)^2 - \frac{1}{4!} - \frac{1}{2}\frac{1}{3!} + \frac{1}{5!} = \frac{2}{15}$$

The probabilities of binary words of length 5, computed with the above algorithm, are shown in table 1. In fig. 1 we compare the probability P_W with the histogram of frequencies in simulated sequences.

4   Tests of independence

The existence of a universal probability distribution for permutations of fixed length enables us to set a framework for a statistical test of independence.

Let $X_1, \ldots, X_N$ be a i.i.d. random sequence and let $\sigma \in S_{n+1}$. We define the estimator

$$T(\sigma) = \left[\frac{N}{n+1}\right]^{-1} \sum_{i=1,\ i\equiv 1 \bmod (n+1)}^{N-n} \chi(\Pi(X_i, \ldots, X_{i+n}) = \sigma) \qquad (3)$$

where: χ is the indicator function which takes the value 1 when the equality $\Pi(X_i, \ldots, X_{i+n}) = \sigma$ is true and 0 otherwise; square brackets denote integer part, hence $\left[\frac{N}{n+1}\right]$ is the number of consecutive disjoint intervals of length $n+1$ in
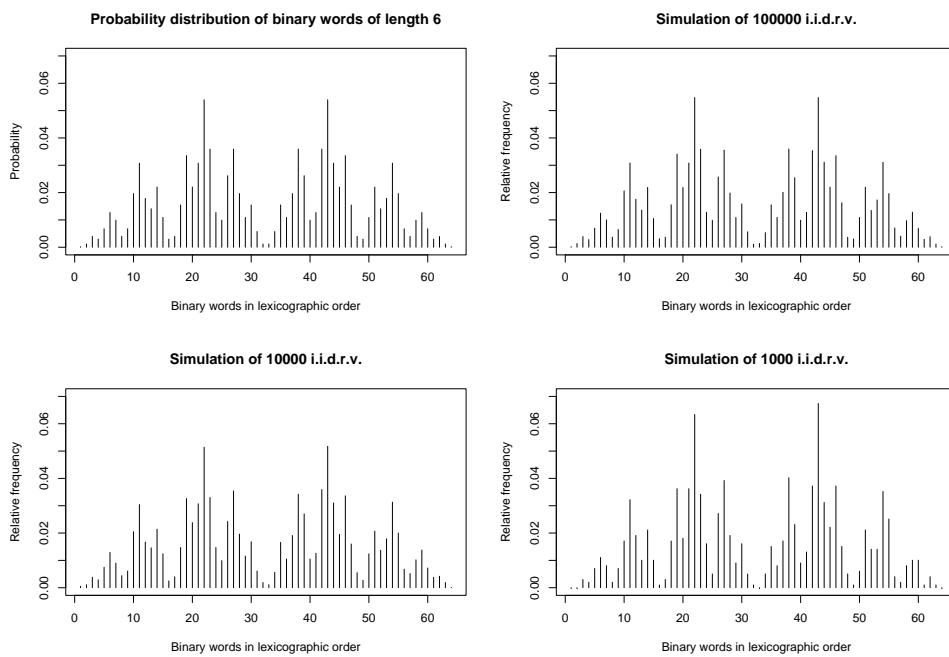
Figure 1: Comparison of the probability distribution of binary words of length 6 and their relative frequency in simulations of a sequence $X_1, \ldots, X_N$ of uniformly distributed and independent random variables for N=100000, 10000, 1000. The main qualitative difference is that for $N = 1000$ the relative frequencies of the most probable words (i.e. 101010 and 010101) are larger than theoretical ones.

which we can split the series; the indexes in the sum are the minimum of these intervals. Let I be the set of these indexes. For brevity we shall write

$$T(\sigma) = \left[\frac{N}{n+1}\right]^{-1} \sum_{i \in I} \chi_i(\sigma).$$

We notice that the random variables $\chi_i(\sigma)$ for $i \in I$ are independent. Hence the $(n+1)!$ variables $T(\sigma)$ form a multinomial vector with $\sum_{\sigma \in S_{n+1}} T(\sigma) = \left[\frac{N}{n+1}\right]$ and with probabilities $P_\Pi(\sigma) = \frac{1}{(n+1)!}$.

If we denote by $\hat{T}(\sigma)$ the value that $T(\sigma)$ takes on the series we assume that the statistics

$$\left[\frac{N}{n+1}\right] \sum_{\sigma \in S_{n+1}} \frac{(\hat{T}(\sigma) - \frac{1}{(n+1)!})^2}{\frac{1}{(n+1)!}} \tag{4}$$

is asymptotically distributed as a $\chi^2((n+1)! - 1)$ (as $N \to \infty$) and we use the standard χ^2 test for goodness of fit (see [9]).

The existence of a universal probability distribution also for words of fixed length enables us to set up a framework for a statistical test of independence for words, even if it appears more difficult to meet the requirements for the chi square test. However words are important from a practical point of view. In fact the number of permutations of length $n+1$ is $(n+1)!$, while the number of binary words of length n is 2^n and therefore it remains manageable for more values of n. Each word corresponds to a subset of permutations and the set of words determines a partition of the set of permutations. We believe that other sets of permutations and other partitions can encode more efficiently than words some specific information about a time series (see for example [2], [3], [5], [7]).

The work we have done for permutations may be essentially repeated for words. Let $X_1, \ldots, X_N$ be a (i.i.d.) random sequence and let $w \in \mathbb{Z}_2^n$. We define the estimator

$$V(w) = \left[\frac{N}{n+1}\right] \sum_{i \in I} \chi(W(X_i, \ldots, X_{i+n}) = w)$$

where χ is the indicator function which takes the value 1 when the equality $W(X_1, \ldots, X_{i+n}) = w$ is true and 0 otherwise. We remark that the set I is the same for permutations of S_{n+1} and words of length n, since two words are independent if and only if their initial indexes differ at least $n+1$ and words are defined in terms of the series of differences. As before we have that the 2^n variables $V(w)$ form a multinomial vector with $\sum_w V(w) = \left[\frac{N}{n+1}\right]$ and probabilities $P_W(w)$ computed in the previous section. The analogue of (4) is

$$\left[\frac{N}{n+1}\right] \sum_{w \in \mathbb{Z}_2^n} \frac{(\hat{V}(w) - P_W(w))^2}{P_W(w)} \tag{5}$$
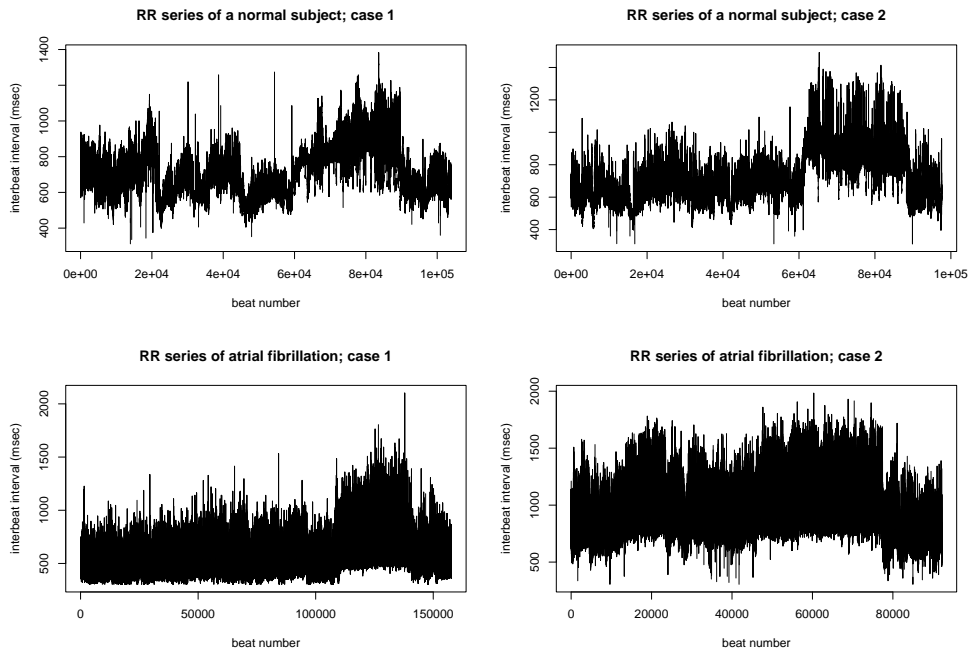
9

Figure 2: Plot of the RR series we have analyzed

5   Data analysis

In this section we analyze the empirical frequencies of permutations and words in the case of *heartbeat series*. The heartbeat series is defined as the sequence of duration of the cardiac cycle measured as the time interval between two consecutive *R* peaks in the ECG recording (RR sequence).

The data we analyze in this paper are the RR sequences extracted from 24 hours Holter recordings of two normal subjects and two subjects with atrial fibrillation. The series were obtained from Holter equipments with sampling frequency of 180*Hz* (Rozinn Electronics, Glendale, USA) and taken from the database provided by the Department of Cardiology of our University. The RR series that we have analyzed are shown in figure 2.

In these time series some of the values are not defined for measurement errors or for other causes. If a segment contains one or more missing values, we shall simply drop it. This implies that in the statistics of formula (4) we need to change the factor $\left[\frac{N}{n+1}\right]$ into the number of segments which do not contain missing data.

We believe that the analysis of $\hat{T}(\sigma)$ and $\hat{V}(w)$ are useful tools for understanding the short range neuroautonomic control mechanisms of heart beat, as explained for example in [4] and [5].

The values of $\hat{T}$ and $\hat{V}$ for the RR series which we have analyzed are sum-
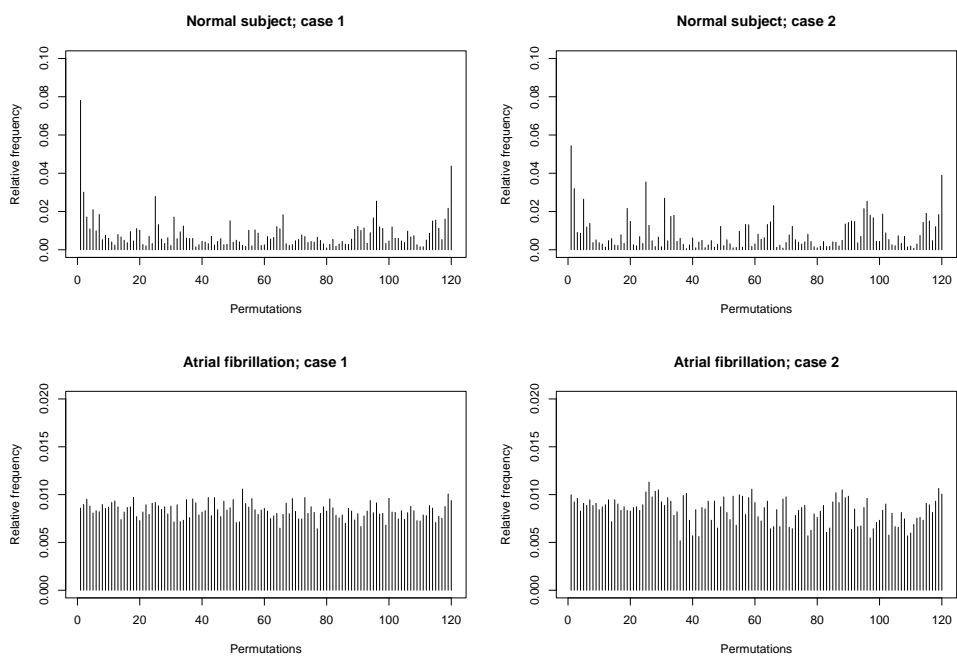
Figure 3: Comparison of the probability distribution of permutations of segments of length 5 and their relative frequency in four RR sequences. Note that the values for the y axes are between 0 and 0.1 in the normal cases and between 0 and 0.02 in the fibrillation ones

11

	nor1	nor2	fib1	fib2	simul i.i.d
permutations	24370	21089	274	362	101
length 5 words	38347	22773	66.8	66.6	18.4
length 4 words	19387	15949	43.8	57.2	9.4

Table 2: Chi square values for binary words and permutations. Only the simulated i.i.d. sequence is below the 95% threshold (145.5 for permutations, 45.0 for binary words of length 5 and 25.0 for binary words of length 4). Note however that the values of chi square is at least two order of magnitude higher for normal than for atrial fibrillation.

marized in the histograms of figures 3 and 4. Even if the third and the fourth histograms of figure 3 seem very close to uniform, the chi square test shows a significant statistical deviation from uniformity. In table 2 we display the values for the χ^2 test for goodness of fit of the frequency of length 5 permutations in our data with respect to P_Π (first row) and of the frequency of length 5 and length 4 words in our data with respect to P_W (second and third row). In the last column we test the fit for a simulated i.i.d. sequence. The 95% confidence threshold is 145.5 for the first row (119 degrees of freedom), 45.0 for the second row (31 degrees of freedom) and 25.0 for the third row (15 degrees of freedom).

By looking at the first row of this table, it is clear that no RR sequence passes the test of independence based on permutations (which is passed by the simulated series). It seems interesting however to use the chi square value as a measure of the distance from independence. This could be one of the possible indexes with which to classify different kinds of atrial fibrillations. The values of chi square are much lower for atrial fibrillation than for normal. This suggests that the RR series of atrial fibrillation is quite close to independence. In the second and third rows of table 2 the values of the chi square statistics for words give results which substantially agree with those for permutations.
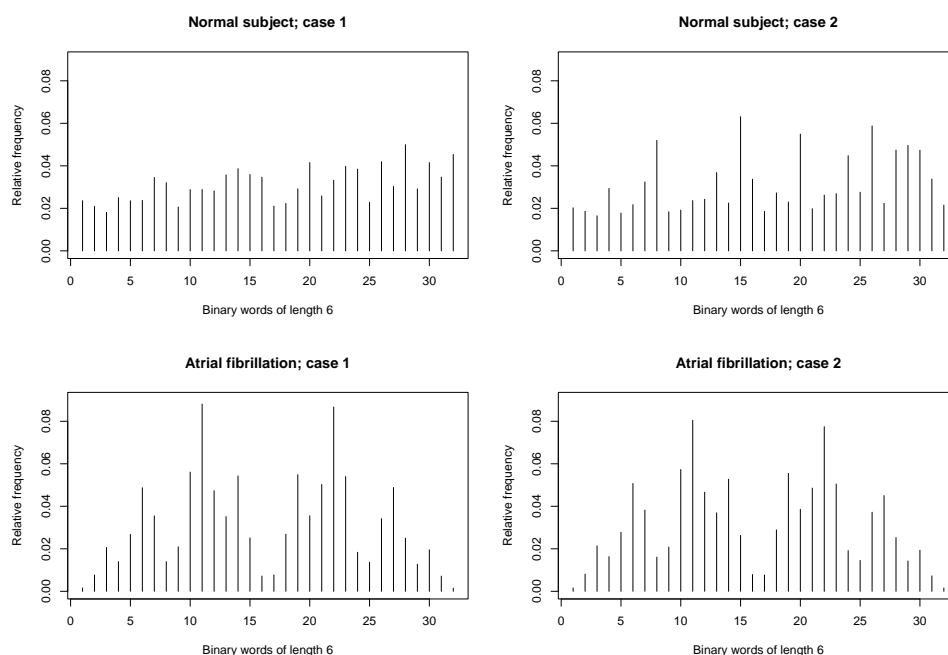
Figure 4: Comparison of the probability distribution of words of length 5 and their relative frequency in four RR sequences

References

[1] Cammarota C. "Estimating correlation exponents of the heartbeat time series" *Int. J. Bifurcation and Chaos*, **10** (2000) , 1513-1520

[2] Ambrosini M., Cammarota C., Guarini G. "Heart beat stationarity in heart transplanted patients", *Clin. Ter.*, **152** (2001), 363-366

[3] Cammarota C., Guarini G., Ambrosini M. "Analysis of stationary periods of heart rate via symbolic dynamics", *Medical Data Analysis*, Lecture Notes in Computer Science vol. 2526, Colosimo, Giuliani and Sirabella eds. Springer 2002

[4] Cammarota C., Guarini G., Rogora E., Ambrosini M. "Non stationary model of the heart beat time series in atrial fibrillation", *Mathematical modeling and computing in Biology and Medicine, Proceedings of 5th ESMTB Conference, Milan 2002* Ed. V. Capasso, Milano, Società Editrice Esculapio 2003

[5] Germanò G., Piccirillo G., Cammarota C., Guarini G., Rogora E., Cacciafesta M. "A new measure of acceleration of heart rate: dependence on age and comparison with time domain conventional heart rate variability

measures", Preprint n. 42/2003 of the Department of Mathematics, University "La Sapienza" of Rome. Available on line at the Internet location `http://w3.uniroma1.it/CISB/Cisb/fr_documents.html`

[6] Albert, C.-C. Yang, Shu-Shya Hseu, Huey-Wen, Yien, Ary L. Goldberger, C.-K. Peng "Linguistic Analysis of the Human Heartbeat Using Frequency and Rank Order Statistics", *Phys. Rev. Lett.*, **90** (2003), 108103-1,108103-4.

[7] Garcia-Gonzales M. A., Ramos-Castro J., Fernandez-Jimeno M. "A new index for the analysis of heart rate variability dynamics: characterization and application", *Physiol. Meas.* **24** (2003), 819-832

[8] Bandt C., Pompe B. "Permutation Entropy: A natural Complexity Measure for Time Series", *Phys. Rev. Lett.* **88** (2002) 174102-1:174102-4

[9] Rohatgi V. K., *An introduction to Probability Theory and Mathematical Statistics*, New York, Wiley ed., 1976

[10] Keller K., Lauffer H., "Symbolic analysis of high dimensional time series", *Int. J. Bifurcation and Chaos* **13** (2003) 2657-2668

[11] Meyerfeldt U. et al. "Heart rate variability before the onset of ventricular tachycardia: differences between slow and fast arrhythmias", *International journal of Cardiology* **84** (2002), 141-151

[12] Zebrowski J.J., Poplawska W., Baranowski R., Buchner T. "Symbolic dynamics and complexity in a physiological time series" *Chaos, solitons and Fractals* **11** (2000) 1061-1075

[13] June-Soo Kim et al. "Decreased entropy of symbolic heart rate dynamics during daily activity as a predictor of positive head up tilt test in patients with alleged neurocardiogenic syncope", *Phys. Med Biol.* **45** (2000) 3403-3412