

Comment on “Linguistic Analysis of the Human Heartbeat Using Frequency and Rank Order Statistics”

In a recent Letter Yang *et al.* [1] propose a “distance” between symbolic sequences, based on rank order statistics of words in these sequences, and use it to analyze heart rate fluctuations. We show here that (i) the quantity they use violates essential properties of a distance measure and (ii) essential results of their Letter, which they claim can be obtained only from their algorithm, can be obtained using simply the average and standard deviation of the heartbeat intervals.

In [1] the authors consider binary sequences and generate the frequency distribution of m -bit words, $p(k)$, with $k = 1, \dots, M (= 2^m)$ being an index number. Using ranking with decreasing probability, $R(k) = 1, \dots, M$ with $R(k) < R(k')$ if $p(k) > p(k')$, they define a distance between sequences [2],

$$D(S_1, S_2) = \frac{\sum_{k=1}^M |R_1(k) - R_2(k)| p_1(k) p_2(k)}{(M-1) \sum_{k=1}^M p_1(k) p_2(k)}, \quad (1)$$

which would actually be a distance between the frequency distributions, i.e., $D(S_1, S_2) = D(p_1, p_2)$. Essential properties of any distance measure d are (i) $d(A, B) = 0$ if and only if $A \equiv B$ [3] and (ii) the triangular inequality

$$d(A, B) + d(B, C) \geq d(A, C). \quad (2)$$

We give counterexamples that violate these properties for D . We start with any distribution p_A that is already rank ordered, $R_A(k) = k$, with the additional property $p_A(M) > 0$. It is easy to see from (1) that any other distribution p_B having the same rank ordering gives $D(p_A, p_B) = 0$, thereby violating the first property. For the following we choose $p_B(M) = 0$ and $p_B(M-1) > 0$. We construct p_C by exchanging the last two ranks; i.e., $p_C(M-1) = p_B(M)$ and $p_C(M) = p_B(M-1)$. Since the corresponding respective probabilities are zero in either p_B or p_C , the distance is zero again, $D(p_B, p_C) = 0$. However, since $p_A(M) > 0$ and $p_C(M) > 0$, and there is a nonzero rank difference, it immediately follows that $D(p_A, p_C) > 0$, thereby violating (2). Any conclusions based on the assumption of (1) being a distance may therefore be erroneous.

While (1) fails as a distance measure it may nevertheless still be useful as a categorizing quantity. Yang *et al.* use it to analyze a binarized derivative of the heartbeat interval time series and compare different groups of individuals: healthy young (Y), healthy elderly (E), patients with congestive heart failure (CHF), and patients with atrial fibrillation (AF). We reanalyzed their data just looking at the mean heartbeat interval and its standard deviation (std) [4]; see Fig. 1. As can be seen, a distinction between the healthy and the different unhealthy groups is possible by just using the mean. Distinction between Y and E becomes possible by using the std. This is

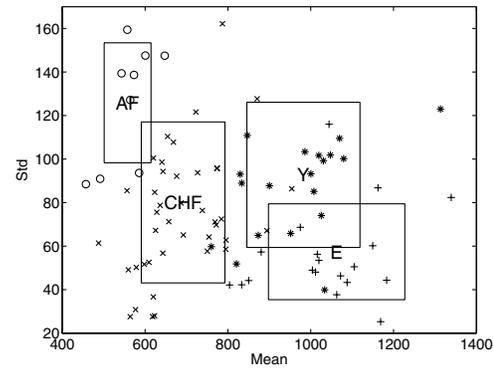


FIG. 1. Mean heartbeat interval and its standard deviation (std) in msec for the data of [1]; (circles) AF, (crosses) CHF, (asterisks) Y, and (plusses) E; the rectangles are defined by the respective group means and group standard deviations.

confirmed by a statistical analysis (all $p < 3 \times 10^{-4}$). Using discriminant analysis employing mean and std we were able to correctly classify 80% of the data sets to their respective groups. Intriguing is the fact that a distinction between healthy and unhealthy groups is possible just using the mean, suggesting that fluctuations are of secondary importance only.

Finally, the authors claim that only their method can effectively discriminate data sets of the AF group from white noise. We find, however, that in most cases the autocorrelation function exhibits a relaxational contribution with both amplitude and correlation time significantly different from zero, thereby clearly exhibiting properties different from white noise.

We thank the authors of [1] for sending us their data.

Alexander Kraskov, Walter Nadler, Harald Stögbauer, and Peter Grassberger

John von Neumann Institute for Computing (NIC),
Forschungszentrum Jülich, D-52425 Jülich, Germany

Received 28 May 2003; published 11 March 2004

DOI: 10.1103/PhysRevLett.92.109801

PACS numbers: 87.19.Hh

- [1] A. C.-C. Yang *et al.*, Phys. Rev. Lett. **90**, 108103 (2003).
- [2] Note that this quantity is actually not well-defined in the cases where the sets of words with nonzero frequency in both distributions do not overlap. Also the ranking is ambiguous if $p(k_1) = p(k_2)$ for $k_1 \neq k_2$. However, both problems could be remedied practically by putting some small noise on the distribution and are not our main point here.
- [3] This property is violated already by the transition from sequences to word frequency distributions: several sequences are mapped onto the same distribution as can be seen immediately for $m = 1$; however, one could argue that this is just a consequence of coarse graining.
- [4] In order to exclude artifacts in the data we disregarded heartbeat intervals larger than 3 sec.