ORDER STATISTIC CORRELATION COEFFICIENT AND ITS APPLICATION TO ASSOCIATION MEASUREMENT OF BIOSIGNALS

*Weichao Xu, Chunqi Chang, Y. S. Hung, S. K. Kwan**

{wcxu,cqchang,yshung,skkwan}@eee.hku.hk
Department of Electrical and Electronic Engineering
The University of Hong Kong
Pokfulam Road, Hong Kong

P. C. W. Fung

hrspfcw@hkucc.hku.hk
Department of Medicine
The University of Hong Kong
Pokfulam Road, Hong Kong

ABSTRACT

In this paper we propose a novel and fast nonlinear association measure based on order statistics and rearrangement inequality. We employ one episode of heart signal, one episode of EEG signal and 1000 white Gaussian noises in our study. Extensive statistical analysis are performed based on one linear model and one nonlinear model. Comparative studies with three other prominent methods are presented. Theoretical derivations and experimental results suggest that our new method has small biasedness, high sensitivity to changes in association, fast computational speed, and robustness under monotone nonlinear transformations.

1. INTRODUCTION

A multitude of methods have been used in the literature of biosignal processing for many years to measure the association between two time series. Among these measures the Pearson's linear correlation coefficient [1–3], Spearman's rho and Kendall's tau [4] are perhaps the most prominent.

The indices mentioned above have many advantages but also some shortcomings. Linear correlation coefficient is very fast, however it will yield misleading results if nonlinearity is involved in the system [5]. On the other hand, the two rank correlation coefficients, Spearman's rho and Kendall's tau, are not as powerful and as fast as Pearson's coefficient when measuring linear associations between biosignals; nevertheless they are independent of increasing nonlinear transformations which makes them suitable for many nonlinear cases [4].

Motivated by the shortcomings of the existing methods, in this paper we propose a new and fast order statistic correlation coefficient which possesses the advantages of all aforementioned methods: a) the new method has comparable performance with Pearson's coefficient when measuring linear association; b) it runs quite fast (in the order of $O(N\log(N))$); and c) it possesses similar properties with the two rank-based coefficients when nonlinearity is involved.

*This work was supported by the Hong Kong Innovation and Technology Commission under Funding ITS/109/02, and in part by Hong Kong RGC grant under N_HKU703/03.

2. THEORY

2.1. Definition of Order Statistics Correlation

Let $(x_i, y_i), i = 1, \ldots, N$ be two time series of length N. Rearranging pairwise the two time series with respect to the magnitudes of x, we get two new series denoted by $(x_{(i)}, y_{[i]})$, where $x_{(1)} \leq \cdots \leq x_{(N)}$ are called the *order statistics* of x and $y_{[1]}, \ldots, y_{[N]}$ the associated *concomitants* [6]. Reversing the roles of x and y, we also define $y_{(1)}, \ldots, y_{(N)}$, and $x_{[1]}, \ldots, x_{[N]}$, respectively. We define our new order statistics based correlation, as follows:

$$r_X(x,y) \triangleq \frac{\sum_{i=1}^{N}(x_{(i)} - x_{(N-i+1)})y_{[i]}}{\sum_{i=1}^{N}(x_{(i)} - x_{(N-i+1)})y_{(i)}} \qquad (1)$$

In the sequel we will use r_P, r_S, r_K, and r_X to denote the Pearson's coefficient, Spearman's rho, Kendall's tau, and our new measure. We may also denote the four measures in general as $r_\xi, \xi = X, P, S, K$.

2.2. Properties of r_X

There are several desirable properties of r_X, as follows:

a) r_X is limited within $[-1, +1]$;

b) $+1(-1)$ are attained when x and y are in monotonic increasing (decreasing) relationship;

c) If x and y are independent identically distributed (IID), the expectation $E\{r_X(x,y)\} = 0$.

Proof. a) According to the rearrangement inequality [7], it follows that:

$$\sum_{i=1}^{N} x_{(N-i+1)}y_{(i)} \leq \sum_{i=1}^{N} x_{(i)}y_{[i]} \leq \sum_{i=1}^{N} x_{(i)}y_{(i)} \qquad (2)$$

and

$$\sum_{i=1}^{N} x_{(N-i+1)}y_{(i)} \leq \sum_{i=1}^{N} x_{(N-i+1)}y_{[i]} \leq \sum_{i=1}^{N} x_{(i)}y_{(i)} \quad (3)$$

Subtracting (2) by (3) and dividing the difference by $\sum (x_{(i)} - x_{(N-i+1)})y_{(i)}$, we have $-1 \leq r_X \leq 1$, whence the result.

b) Assume $y_i = \phi(x_i), i = 1, 2, \ldots, N$. If $\phi(\cdot)$ is an increasing function, we have $y_{[i]} = y_{(i)}$ for all i. Substitute this into (1), we have $r_X = 1$. Similarly $r_X = -1$ if $\phi(\cdot)$ is a decreasing function.

c) Denote the numerator and denominator of (1) by U and V, respectively. Using a Taylor series expansion of U/V around the expectations $E(U)$, $E(V)$ and ignoring all terms of order higher than two, we have [8]

$$E(r_X) \approx \frac{E(U)}{E(V)} + \text{var}(V)\frac{E(U)}{E^3(V)} - \frac{\text{cov}(U,V)}{E^2(V)} \quad (4)$$

Imposing the independence assumption, we have

$$E(U) = \sum [E(x_{(i)}) - E(x_{(N-i+1)})]E(y_{[i]}) \quad (5)$$

Denote the probability density function (PDF) of $y_{[i]}$ by $g_{[i]}(y)$. It is known [9] that

$$g_{[i]}(y) = \int_{-\infty}^{+\infty} f(y|x)f_{(i)}(x)dx \quad (6)$$

where $f(y|x)$ is the conditional density function of y given x and $f_{(i)}(x)$ the PDF of $x_{(i)}$. The conditional density function $f(y|x)$ degenerates to $f(y)$ if x and y are independent. Then we have

$$\begin{aligned} E(y_{[i]}) &= \int y g_{[i]}(y)dy = \int yf(y)dy \int f_{(i)}(x)dx \\ &= \int yf(y)dy = E(y) \end{aligned} \quad (7)$$

Substituting (7) into (5), we have $E(U) = 0$, and therefore $E(U)E(V) = 0$. On the other hand, via straightforward algebra, we get

$$\begin{aligned} E(UV) = &\sum\sum E[x_{(i)}x_{(j)}]E[y_{[i]}y_{(j)}] \\ &- \sum\sum E[x_{(i)}x_{(j)}]E[y_{[i]}y_{(N-j+1)}] \\ &- \sum\sum E[x_{(N-i+1)}x_{(j)}]E[y_{[i]}y_{(j)}] \\ &+ \sum\sum E[x_{(N-i+1)}x_{(j)}]E[y_{[i]}y_{(N-j+1)}] \end{aligned} \quad (8)$$

Observing that the probability of $y_{[i]} = y_{(j)}$ equals $1/N$, and using the IID assumption, we have $E[y_{[i]}y_{(j)}] = \frac{1}{N}E(y^2)$, hence the first item in (8) becomes $\frac{1}{N}E(y^2)E[(\sum x_i)^2]$. Similar argument holds for the other three items in (8), which means $E(UV)$ is also 0, thus proves that $\text{cov}(U,V) = 0$. Substituting these facts into (4), we have $E(r_X) = 0$, which completes the proof.

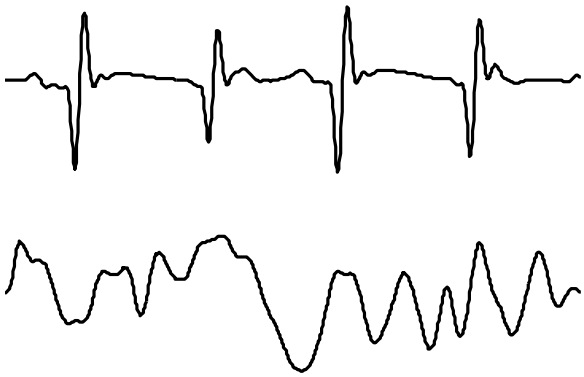

Fig. 1. The heart signal and EEG signal used in the study. The upper one is a smoothed version of a flutter signal, the lower one is a smoothed version of an EEG signal.

3. METHODS

3.1. Signals Employed in This Study

Biosignals can be visually classified into two categories: a) spiky semi-periodic signals (such as normal ECG) composing sharp pulses rested on a flat baseline, and b) signals exhibiting noise-like patterns (such as EEG or Atrial Fibrillation recordings). We employ two episodes of real signals for the purpose of comparison. One is a second of bipolar intra-atrial flutter signal recorded and sampled at 1000Hz during electrophysiological procedure [10], the other is a second of EEG signal (sampling rate 256Hz) from a dataset provided by University of Tuebingen for BCI Competition 2003 [11]. The EEG signal was re-sampled at 1000Hz for the aim of consistency. The two real signals were further filtered respectively by two low-pass filters. The resulting two signals are denoted by s_h and s_e, respectively and shown in Fig. 1.

We generated 1000 episodes of independent white Gaussian noises ($\mu = 0$ and $\sigma^2 = 1$) for the aim of statistical analysis. All the simulated signals contain 1000 samples ($N = 1000$). Without loss of generality, $s_\zeta, \zeta = h, e$ were standardized with mean zero and variance unity before feeding them into the two models described in the following subsection.

3.2. Linear and Nonlinear Models

In this paper we employ one linear model and one nonlinear model for comparative studies and denote them respectively by LM and NM. Under each model, two channels of signals x and y are generated with $s_\zeta, \zeta = h, e$ and 1000 episodes of white noises described in the last subsection. Four sets of correlation coefficients between x and y are then computed for further comparative study. Due to the 1000 noises involved, each r_ξ becomes a random variate and has a distribution, which allows us to perform statistical analysis. In both

of the following models, time index i is from 1 to 1000, $n(i)$ denotes the noises, $\bar{r}_\xi$ and v_ξ denote the mean and standard deviation of $\bar{r}_\xi$, respectively. The Linear Model is a regression model of the form

$$\begin{aligned} x(i) &= s(i) \\ y(i) &= \rho s(i) + \sqrt{1-\rho^2}n(i) \end{aligned} \tag{9}$$

where $\rho = \{-1, -0.99, \ldots, 0.99, 1\}$ characterizes the linear association. It follows from straightforward algebra that Pearson's coefficient r_P is unbiased and hence $E[r_P(x,y)] = \rho$ for any distribution of $s(i)$. Unfortunately this desirable property does not hold for the other three coefficients under this model. The aim of this model is to compare the biasedness of these three biased estimators as well as their power to discriminate different ρ's. The nonlinear model used to study the effect of nonlinar transformations to the four coefficients is as follows:

$$\begin{aligned} x(i) &= T_x[\beta \cdot s(i)] \\ y(i) &= T_y[\beta\{\rho s(i) + \sqrt{1-\rho^2}n(i)\}] \end{aligned} \tag{10}$$

where, in this study, $T_x[\cdot] = \mathrm{sgn}(\cdot)(\cdot)^2$ and $T_y[\cdot] = \exp(\cdot)$. The parameter β is to control the extent of nonlinearity (greater value of β corresponds to stronger nonlinearity), while ρ has the same meaning as in LM.

3.3. Performance Evaluation

Several methods were used to evaluate the performance of r_ξ under both of the two models mentioned above.

3.3.1. Sensitivity to Changes in ρ

Given two distinct ρ_1 and ρ_2 ($\rho_2 > \rho_1$), we have two sets of coefficients $r_{\xi 1}$ and $r_{\xi 2}$, and their respective Fisher's z Transformation [1], $z_{\xi 1}$ and $z_{\xi 2}$. The *Sensitivity ratio* (SR) [3] is defined as:

$$SR_\xi = \frac{\bar{z}_{\xi 2} - \bar{z}_{\xi 1}}{\sqrt{v_{z1}^2 + v_{z2}^2}} \tag{11}$$

where v_{z1} and v_{z2} denote respectively the standard deviation of $z_{\xi 1}$ and $z_{\xi 2}$. SR measures the ability of r_ξ to detect the changes of the underlying ρ. Greater value of SR indicates better discrimination in sensitivity. SR is performed on the resultants of linear model LM and the nonlinear model NM.

3.3.2. Time Complexity Measurement

We analyzed the time complexities of r_ξ in the language of *big Oh*, which is popular in algorithm analysis. We also estimated the relationship between computational loads of r_ξ versus the length of signal N from 100 to 1000 with $\Delta N = 100$.

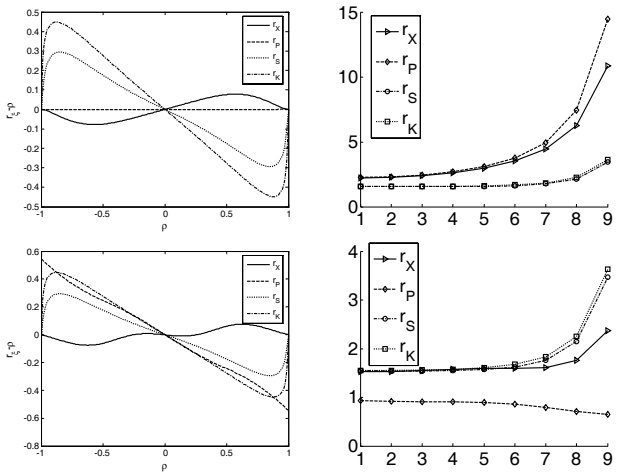

Fig. 2. The performance of heart signal under the two models. The left column shows $r_\xi - \rho$ vs. ρ; the right column shows the SR of the four measures. The upper row corresponds to the performance under the linear model; the lower row to the performance under the nonlinear model. It is noteworthy that in the left-hand panels, the nearer a curve to the zero line, the smaller the corresponding biasedness.

4. RESULTS

Fig.2 and Fig.3 show respectively the results of s_h and s_e under the two models. The left columns of the Fig.2 and Fig.3 illustrate the biasedness of the four measures under the two models, whereas the right columns depict the corresponding sensitivity ratios. It is clear that the performance can be ordered as $SR_P > SR_X > SR_K > SR_S$ under the linear model. On the other hand, the performance can be ordered as $SR_K > SR_S > SR_X > SR_P$ under the nonlinear model ($\beta = 2$). The fastest method is Pearson's coefficient r_P having a linear time complexity of $O(N)$. Our new method r_X and Spearman's rho are of the same order $O(N\log(N))$, since sorting operation dominants the computational time of both methods. However, because of the extra procedure of ranking involved in calculation of r_S , we can expect that r_X is a little faster than r_S. Kendall's tau r_K is the slowest method compared to other three coefficients. The core operation of r_K is to calculate the number of concordant and discordant pairs, which requires $N(N-1)/2$ operations. Therefore, the time complexity of r_K is of order $O(N^2)$.

To confirm this result we estimated the relationship between computational loads of r_ξ versus the length of signal N, where N begins at 100 and increases by 100 successively till $N = 1000$. All the computational speed tests were performed in Matlab 7.0 environment on a Pentium powered PC. For each pair of time series of size N, the algorithms for computing r_ξ were run for 1000 times. The results are presented in Fig.4, which are consistent with our analysis.

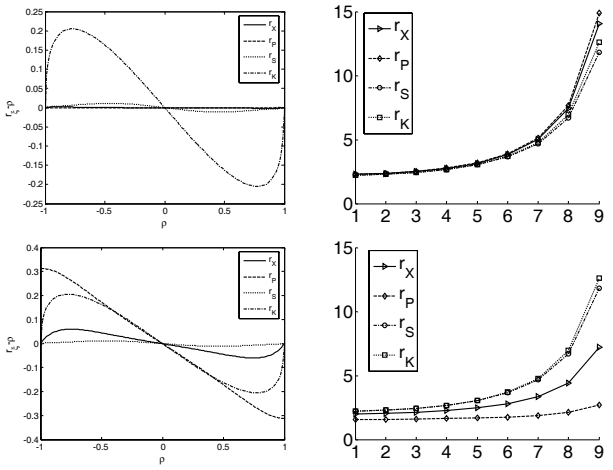

Fig. 3. The performance of EEG signal under the two models. The left column shows $r_\xi - \rho$ vs. ρ; the right column shows the SR of the four measures. The upper row corresponds to the performance under the linear model; the lower row to the performance under the nonlinear model. It is noteworthy that in the left-hand panels, the nearer a curve to the zero line, the smaller the corresponding biasedness.

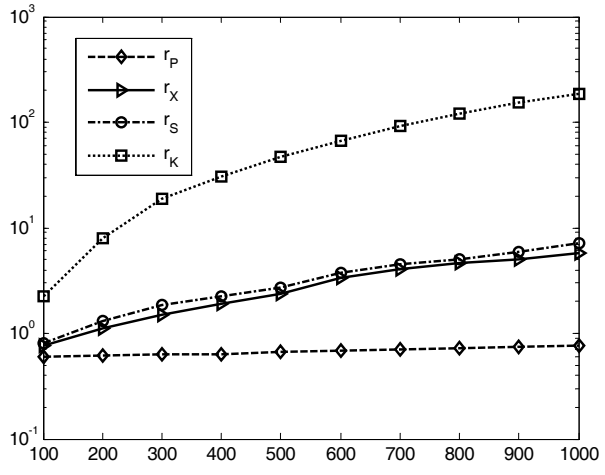

Fig. 4. Results of comparative CPU time test for four coefficients studied. A logarithmic scale is used for better visual effect.

5. CONCLUSION

Our new measure r_X appears to play the role of a "missing link" between Pearson's coefficient and Spearman's *rho* and Kendall's *tau*. It enjoys some advantages of all the three method employed in our comparative studies. In most cases, r_X is not optimal, but it usually is the second best compared to r_P, r_S, and r_K. This feature at least avoids the worst results in practice when one has no prior knowledge whether nonlinearity exists in the system.

6. REFERENCES

[1] R. A. Fisher, *Statistical Methods, Experimental Design, and Scientific Inference*, Oxford University Press, New York, 1990.

[2] R. A. Fisher, "On the 'probable error' of a coefficient of correlation deduced from a small sample," *Metron*, vol. 1, pp. 3–32, 1921.

[3] E. C. Fieller, H. O. Hartley, and E. S. Pearson, "Test for rank correlation coefficients.i," *Biometrika*, vol. 44, pp. 470–481, 1957.

[4] M. Kendall and J. D. Gibbons, *Rank Correlation Methods*, Oxford University Press, New York, 5th edition, 1990.

[5] J. P. S. Cunha and P. G. de Oliveira, "A new and fast nonlinear method for association analysis of biosignals," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 6, pp. 757–763, 2000.

[6] H. A. David and H. N. Nagaraja, *Order Statistics*, Wiley-Interscience, Hoboken, 3rd edition, 2003.

[7] D. S. Mitrinovic, J. E. Pecaric, and A. M. Fink, *Classical and New Inequalities in Analysis*, Kluwer Academic, Dordrecht, 1993.

[8] G. M. P. van Kenpen and L. J. van Vliet, "Mean and variance of ratio estimators used in fluorescence ratio imaging," *Cytometry*, vol. 39, pp. 300–305, 2000.

[9] D. D. Mari and S. Kotz, *Correlation and Dependence*, Imperial College Press, London, 2001.

[10] W. Xu, H. F. Tse, P. C. W. Fung, F. H. Y. Chan, K. L. F. Lee, and C. P. Lau, "New Bayesian discriminator for detection of atrial tachyarrhythmias," *Circulation*, vol. 105, pp. 1472–1479, 2002.

[11] V. Bostanov, "Bci competition 2003-data sets ib and iib: feature extraction from event-related brain potentials with the continuous wavelet transform and the t-value scalogram," *IEEE Transactions on Biomedical Engineering*, vol. 51, pp. 1057–1061, 2004.