New Correlation Coefficient as a Novel Association Measurement With Applications to Biosignal Analysis

Hai Peng, Michael Shell

***Abstract*—AbstractIn this paper,we propose a new correlation coefficient based on Pearsons coefficient and order statistics. Theoretical derivations show that our new coefficient processes the same basic properties as the Pearsons coefficient. Experimental studies based on four models and five biosignals show that our new coefficient performs better than Pearsons coefficient when measuring monotone nonlinear associations. Extensive statistical analyses also suggest that our new coefficient has superior anti-noise robustness, small biasedness, accurate time-delay detection ability, fast and robustness under monotone nonlinear transformations.**

***Index Terms*—new correlation coefficient,Pearson's product moment correlation coefficient(PPMCC),nonlinear association measure,rearangement inequality,order statics correlation coefficient(OSCC).**

I. Introduction

CORRELATION coefficient is a coefficient that illustrates a quantitative measure of some type of correlation and dependence, meaning statistical relationships between two random variables or observed data values.It is important and useful for research prevailing in many scientific and engineering areas,not to mention the area of signal processing.As a measure of such strength,the coefficient should be large and positive if there is a high probability that large(small) values of one time series are associated with large(small) value of another.On the other hand,it should be large and negative if the direction is inverse,namely,large(small) values of one time series occur in conjunction with small(large) values of another . As we known,Pearsons product moment correlation coefficient (PPMCC) [2][4] is a most widely used correlation coefficient for indicating linear associations.However,it will underestimate the strength of association if nonlinearity is involved in the system.On the other hand,the two rank correlation coefficient,Spearmans rho(SR) and Kendalls tau, are suitable for nonlinear cases but they are not as powerful and as fast as Pearsons coefficient for linear cases. Recently,a novel correlation coefficient called order statistics correlation coefficient(OSCC) was proposed.OSCC has similar properties with PPMCC when measuring linear association and possesses comparable performance with the two rank coefficient.In this paper,we propose another method that perform a little better than OSCC. In Section II,we gives the definition and properties of our new correlation coefficient.Section III depicts three linear and one nonlinear models we use in this study.In Section IV,we present the simulated signals and the associated results under four models used in our investigation.Finally,Section V draws the conclusions on the new correlation coefficient.

II. New Correlation Coefficient

A. Definition and Properties

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

$$r_N(x,y) \triangleq \begin{cases} \dfrac{r_P(x,y)}{r_P(x',y')} & \text{if } r_P \geq 0 \\ \dfrac{r_P(x,y)}{r_P(z',y')} & \text{if } r_P < 0 \end{cases}, \tag{1}$$

where:

- r_p is the Pearson's coefficient of (x, y);
- (x', y') is the order statistics of x and y;
- let $z_i = -x_i$,z' is the order statistics of z.

For a sample,the formula for r_P is:

$$r_P = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean);and analogously for $\bar{y}$. Replacing r_P by (2) in (1) gives us this formula for r_N:

$$r_N = \begin{cases} \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_{(i)} - \bar{x})(y_{(i)} - \bar{y})}, & \text{if } r_P \geq 0 \\ \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_{(i)} - \bar{z})(y_{(i)} - \bar{y})}, & \text{if } r_P < 0. \end{cases} \tag{3}$$

*Theorem 1:*The new correlation coefficient has the basic properties of a correlation coefficient,as follows:

1) $-1 \leq r_N \leq 1$;
2) $r_N(x, y)$attains $+1(-1)$ when x and y are in strict increasing(decreasing) relationship;
3) $r_N(x', y') = r_N(x, y)$ for $x' = k_x x + const_x$ and $y' = k_y y + const_y$,where $k_x > 0$ and $k_y > 0$;
4) if x and y are mutually independent and each is independent identically distributed(IID),the expectation $E\{r_N(x, y) = 0\}$ when $N \to \infty$.

Proof:

1) When $r_P \geq 0$, according to the rearrangement inequality,it follows that:

$$0 \leq \sum_{n}^{i=1}(x_i - \bar{x})(y_i - \bar{y}) \leq \sum_{n}^{i=1}(x_{(i)} - \bar{x})(y_{(i)} - \bar{y}) \quad (4)$$

Dividing the 4 by $\sum_{n}^{i=1}(x_{(i)} - \bar{x})(y_{(i)} - \bar{y})$,we have $0 \leq r_N \leq 1$; When $r_P < 0$,we know that $z_{(i)} = -x_{n-i+1}$ and $\bar{z} = -\bar{x}$ according to definition.Then we have

$$\begin{aligned} &-\sum_{n}^{i=1}(z_{(i)} - \bar{z})(y_{(i)} - \bar{y}) \\ =&\sum_{n}^{i=1}(x_{n-i+1} - \bar{x})(y_i - \bar{y}) \\ \leq&\sum_{n}^{i=1}(x_{(i)} - \bar{x})(y_{(i)} - \bar{y}) \end{aligned} \quad (5)$$

2) Assume $y_i = \phi(x_i), i = 1, \dot{,}N$.If $\phi(\bullet)$ is astrict increasing function,we have $r_P > 0$ and $\sum_{n}^{i=1}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{n}^{i=1}(x_{(i)} - \bar{x})(y_{(i)} - \bar{y})$.Substituting this into 3,we have $r_N = 1$;and similarly $r_N = -1$ if $\phi(\bullet)$ is astrict decreasing function.
3) Substituting x' and y' into 3,when $r_P(x', y') \geq 0$,we have

$$\begin{aligned} r_N(x', y') =&\frac{\sum({x_i}' - \bar{x}')({y_i}' - \bar{y}')}{\sum({x_{(i)}}' - \bar{x}')({y_{(i)}}' - \bar{y}')} \\ =&\frac{k_x k_y \sum(x_i - \bar{x})(y_i - \bar{y})}{k_x k_y \sum(x_{(i)} - \bar{x})(y_{(i)} - \bar{y})} \\ =&r_N(x, y); \end{aligned} \quad (6)$$

when $r_P(x', y') < 0$,we have

$$\begin{aligned} r_N(x', y') =& \frac{\sum({x_i}' - \bar{x}')({y_i}' - \bar{y}')}{\sum({z_{(i)}}' - \bar{z}')({y_{(i)}}' - \bar{y}')} \\ =& \frac{k_x k_y \sum(x_i - \bar{x})(y_i - \bar{y})}{k_x k_y \sum(z_{(i)} - \bar{z})(y_{(i)} - \bar{y})} \\ =& r_N(x, y); \end{aligned} \quad (7)$$

Hence,we have $r_N(x', y') = r_N(x, y)$.
4) Denote the numerator and denominator of 1 by U and V,respectively.An application of the Delta method yields

$$Er_N(x, y) = \frac{E(U)}{E(V)} + O(N^{-1}). \quad (8)$$

According to 1,we know that $E(U) = Er_P(x, y) = 0$ where x and y are IID.Hence,we have $E(r_N) = 0$ to the order of $O(N^{-1})$.

B. Estimation of Correlation Coefficient in Normal Case

As we known,the expectation and variance for Pearson's correlation coefficient(r_P) of a normal bivariate are respectively $E(r_P) = \rho$ and $Var(r_P) = \frac{(1-\rho^2)^2}{n-1}$.It is clear that $r_P x_s, y_s \to 1$ because x_s and y_s are identically distributed and increasing.SO we can suppose that $E(r_N) = \rho$ according to Delta method.We employ two channels of signals (x, y) $N(0, 1, \rho)$ to calculate the expection and variance for new correlation coefficient(r_N).The result compared with Pearson'correlation coefficient(r_P) show n in Fig.It is easily observed that the expection and variance for new correlation coefficient(r_N) are similar to those for r_P.

III. Models of Association and Performance Evaluation

In this section,we introduce three linear models and on nonlinear model to model the linear and nonlinear association between two time series.In each model,a time series $x(i)$ is derived from a pure signal $s(i)$,and another signal $y(i)$ is obtained as a combination of the transformed pure signal and a white noise,$n(i)$.In all these models,the time index i is runs from 1 to 100.Models if association are as follow:

*1) Linear Model 1(LM1):*LM1 is construsted as

$$\begin{aligned} x(i) &= s(i) \\ y(i) &= s(i) + \alpha \cdot n(i) \end{aligned} \quad (9)$$

where $\alpha \in [0, 1]$ is increased from 0 to 1 with a step $\Delta\alpha = 0.1$ to control the signal-to-noise ratio (SNR).With increasing α,the association between x and y becomes smaller and smaller,which means tha $r_\xi(\rho_\xi)$ should have a decreasing relationship with α.For a fixed α ,the greater the magnitude of $E(\rho_\xi)$,the better its performance in the context of noise robustness.

*2) Linear Model 2(LM2):*LM2 is a regression model of the form(cite)

$$\begin{aligned} x(i) &= s(i) \\ y(i) &= \rho \cdot s(i) + \sqrt{1-\rho^2} \cdot n(i) \end{aligned} \quad (10)$$

where $\rho \in [-1, 1]$ with a step $\Delta\rho = 0.01$ characterizing the linear association.It follows by straightforward calculation that $E(\rho_P)$ for any distribution of $s(i)$.

*3) Linear Model 3(LM3):*LM3 is similar to LM1 except for a time delay $\delta = 30$ introduced in channel y,as follows:

$$\begin{aligned} x(i) &= s(i) \\ y(i) &= s(i - \delta) + \alpha \cdot n(i) \end{aligned} \quad (11)$$

*4) Nonlinear Model(NM):*NM is a nonlinear model used to study the effect of nonlinear transformations to the signals on the two coefficients,as follows(cite):

$$\begin{aligned} x(i) &= T_x[\beta \cdot s(i)] \\ y(i) &= T_y\left[\beta\left\{\rho \cdot s(i) + \sqrt{1-\rho^2} \cdot n(i)\right\}\right] \end{aligned} \quad (12)$$

where $T_x[\bullet]$ and $T_y[\bullet]$ are two increasing nonlinear functions.The parameter $\beta = 2, 4, 6, 8, 10$ is used to control the extent of nonlinearity (greater value of β corresponding to stronger nonlinearity),while ρ has the same meaning as in LM2.

IV. Comparison on Results for Simulated and Real Biosignals

With the purpose of evaluating the feasibility of our new correlation coefficient in association studies,five main types of biosignals are employed for investigation.They

are periodic,semi-periodic,stationary,nonstationary,and long-range-correlated signals which are denoted as $s_\zeta, \zeta = p, h, a, e, l$ for notational convenience.Whats more,periodic and semi-periodic signals are classified as deterministic signals which can be described by explicit mathematical relationship;whereas stationary and nonstationary signals are classified as stochastic signals which can be described only in statistical terms.The last one of five types of biosignals is the biosignal that exhibit long range power-law correlation.A stochastic process with long range correlation means that its autocorrelation function $R(k)$ k^{2H-2} as $k \to \infty$,where $0 < H < 1$ is the Hurst parameter.Its corresponding power spectral density is proportional to $f^{-(2H-1)}$.In addition,1000 episodes of independent white Gaussian noise($\mu = 0$ and $\sigma^2 = 1$) are generated to do duty for noises which are involved in the linear and nonlinear models.Therefore,we can perform statistical analysis because each r_ξ becomes a random a random variable and has a distribution.

A. Simulated and Real Biosignals

As remarked before,the following five representative biosignals are included in our study:

1) sin wave $s_P(i)$ of frequency 5 Hz emulating periodic biosignals;
2) real bipolar intra-atrial flutter signal $s_h(i)$ recorded during electrophysiological procedure(cite);
3) episode of alpha wave $s_a(i)$ simulated from a random Gaussian noise filtered by a band-pass Butterworth filter with passband 8 to 12 Hz (cite);
4) second of real EEG signal $s_e(i)$ (sampling rate 256 Hz) from a dataset provided by University of Tuebingen for BCI Competition 2003 (cite);
5) segmentation $s_l(i)$ of an artificial time series $s_{lf}(i)$ exhibiting long range correlation with Hurst parameter $H = 0.9$ (cite).

Fig.2 illustrates the five biosignals above.All the first four signals among them possess one thousand samples.The EEG signal $s_e(i)$ is up-sampled from 256 to 1000 Hz by linear interpolation.As for $s_l(i)$,we choice the 1000 samples of $s_{lf}(i)$ for association analysis.Before feeding s_ζ into the linear and nonlinear models,we normalized them to have mean zero and variance unity.

B. Comparative Study Under Linear Model LM1

Fig.3 illustrates the result that $\rho_\xi(\alpha)$ drops with increasing of α under LM1.In Fig.3a-3b,the decreasing rate of ρ_N and ρ_X are more slow than that of ρ_P.It means that the new coefficient and OSCC perform superiorly when deterministic signals s_P and sh are fed into LM1.Whats more,ρ_N even outperforms ρ_X for s_h.On the other hand,there is little differences observed in Fig.3c-3e when the inputs are stochastic signals s_a s_e and s_l.From the above,the new coefficient has a better noise robustness performance than OSCC and PPMCC.

C. Comparative Study Under Linear Model LM2

The relationships between $\bar{\rho_\xi}$ and ρ for the five biosignals s_ζ are shown in Fig.4.It is clear that 1)$-1 \leq \rho_\xi$(Property 1);2)$\rho_\xi = \pm 1 (r_\xi = \pm 1)$ as $\rho = \pm 1$,respectively(Property 2);3) $\bar{\rho_\xi} = 0 (\bar{r_\xi} = 0)$ as $\rho = 0$(Property 4);and 4)$\bar{\rho_\xi}(\bar{\rho_\xi})$ is an increasing function of ρ.For the reason that the closer the distance of $\bar{\rho_\xi}$ to the diagonal line,the smaller the associated biasedness,the unbiasedness performance can be ordered as $r_P > r_N, rX$ when deterministic signals s_h is the model inputs;as for order four signals,there are immaterial differences among the two methods.

D. Comparative Study Under Linear Model LM3

Under this model,r_ξ is computed as a function of time-shift κ,say,which varies from -100 to 100 ms.For each α and each episode n_i of 1000 white noises,$r_\xi(\alpha, \kappa)$ is calculated and the time-shift with respect to the maximum of $r_\xi(\alpha, \kappa)$ is the estimate of the time-delay Δ and denoted by κ_Δ.Limited by the length of this paper,we only present the results with respect to s_e here.Fig.5a shows the waveforms of $r_\xi(\alpha, \kappa)$ in the presence of a 50% SNR($\alpha = 1$).All the coefficients can correctly detect the time-delay between x and y giving $\kappa_\Delta = 30ms$ which equals the true time-delay Δ.In Fig.5b,we present the statistical results of κ_Δ versus the underlying α from 0 to 1 with $\Delta\alpha = 0.1$.The levels of rectangular bars represent the means $\bar{\kappa_\Delta}$ and the error bars represent $3 \times \upsilon_{\kappa_\Delta}$ with υ_{κ_Δ} denoting the standard deviation of κ_Δ.From Fig.5b it can be found that $\bar{\kappa_\Delta}$ slightly increases with increase of noise levels and so does the standard deviation υ_{κ_Δ} for all r_ξ.The performance of time-delay detection is so unconspicuous that we do not think that there are significant difference between those methods in the aspect of detecting time delays.

E. Comparative Study Under Nonlinear Model NM

The nonlinear model NM is constructed on the linear model LM2 by introducing two increasing nonlinear transformations $T_x[\bullet] = sgn(\bullet)(\bullet)^2$ and $T_y[\bullet] = exp(\bullet)$.In addition to association parameter ρ playing the same effect as in LM2,another parameter ρ is employed to control the extent of nonlinearity.

Giving nonlinearity parameter $\beta = 2$,the relationships between $\bar{\rho_\xi}$ and ρ are shown in Fig.6. It can be observe that ρ_N and ρ_X have smaller biasedness than ρ_P.Moreover,r_P never approaches ± 1 as $\rho \to \pm 1$ while $r_N = \pm 1, r_X = \pm 1$ as $\rho \to \pm 1$.It means that r_P underestimates the strength of association when nonlinearity is involved but r_N and r_X have superiority under increasing nonlinear transforms.

V. Conclusion

In this paper,we propose a new correlation coefficient and investigate its properties.The proposed measure was evaluated using simulated and real biosignals and four models emulating linear and nonlinear situations.We also compared the behavior of our measure with PPMCC and OSCC.The comparative studies demonstrate that our new correlation coefficient performs well in whether linear or nonlinear cases and it enjoys the advantages of the PPMCC and OSCC.In most cases,the new correlation coefficient is not optimal,but it usually is the great substitution.This suboptimal feature at least avoids the worst results in practice when one has no prior knowledge as to whether the system is nonlinear

REFERENCES