Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physica D 200 (2005) 287–295
PHYSICA D
www.elsevier.com/locate/physd

A nonlinear correlation measure for multivariable data set

Qiang Wang[a,*], Yi Shen[a], Jian Qiu Zhang[b]

[a] *Department of Control Science and Engineering, Harbin Institute of Technology, PO Box 327, 92# XiDaZhi Street, Harbin 150001, PR China*
[b] *Department of Electronic Engineering, Fudan University, Shanghai, PR China*

Received 22 November 2003; received in revised form 14 July 2004; accepted 9 November 2004

Communicated by S. Kai

Abstract

The paper states that the mutual information carried by the rank sequences that are obtained from the original two sequences is a good measure of nonlinear correlation. Based on that, the nonlinear correlation information entropy (NCIE) is proposed for measuring the general relationship of a multivariable data set. NCIE uses a number in the closed interval $[0, 1]$ to indicate the nonlinear correlation degree of the concerned multivariable data set, with 0 and 1 denotes the weakest and the strongest relationship, respectively. Data sets generated from an autoregressive model, a nonlinear chaotic Lorenz system, and a logistic function as known correlation sequences are analyzed using the proposed NCIE. The results testify the suitableness and correctness of the proposed concepts as nonlinear correlation measure.

PACS: 05.70; 05.45; 89.70; 71.45.G

Keywords: Entropy; Nonlinear correlation; Mutual information; Correlation coefficient

1. Introduction

In Shaw's essay, chaotic dynamical systems were interpreted as generators of information [1]. This information-theoretic viewpoint has led to a number of algorithms and statistics which characterize strange and chaotic attractors in terms of their information-generating properties. Foremost among these is the Kolmogorov–Sinai (KS) entropy [1,2], which measures the average production of Shannon [3] entropy per unit time. Generalizations of this quantity have been described [4–6] which measure the production rate of Renyi [7] entropies. Various measures of dependence have also been employed, such as mutual information [1,8] (between two variables) and redundancy [9,10] (among several variables). Another commonly used information-theoretic statistic is the information dimension [11] which

* Corresponding author. Tel.: +86 451 86413411 8602; fax: +86 451 86418378.
E-mail address: wangqiang@hit.edu.cn (Q. Wang).

0167-2789/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.physd.2004.11.001

describes how information in a strange attractor scales with characteristic coarse grain size. These information-theoretic measures are not the only tools available for characterizing nonlinear systems. Measures for characterizing chaos are usually categorized as either "static" or "dynamical", though this distinction is actually somewhat artificial for time series applications, since time-delay embeddings effectively encode dynamical information directly into the static attractor [12].

In this paper, we first demonstrate that the mutual information carried by the rank sequences, which are obtained from the original sequences, is a good measure of nonlinear correlation. Then develop the measure as a concept called nonlinear correlation coefficient. Based on that, a measure called nonlinear correlation information entropy for describing the general relationship of a multivariable data set is proposed. The proposed measure uses a number in the closed interval $[0, 1]$ to indicate the nonlinear correlation degree of the concerned multivariable data set, with 0 and 1 denotes the weakest and the strongest relationship, respectively. Examples of utilizing the proposed concepts to analyze data sets generated by a linear autoregressive function, a nonlinear chaotic Lorenz system and a logistic function are also presented. The conclusions are given at the end of this paper.

2. Nonlinear correlation information entropy

For describing the general correlation between two variables except for the correlation coefficient which is used to describe the linear correlation of the two variables, the mutual information concept is used widely which is defined as

$$I(X; Y) = H(X) + H(Y) - H(X, Y), \tag{1}$$

where X, Y are two discrete random variables concerned. $H(X)$ is the information entropy of the variable X, which is defined as

$$H(X) = -\sum_{i=1}^{L} p_i \ln p_i, \tag{2}$$

and the joint entropy of the two variables X and Y, $H(X, Y)$, is defined as

$$H(X, Y) = -\sum_{i=1}^{L}\sum_{j=1}^{M} p_{ij} \ln p_{ij}. \tag{3}$$

Mutual information can be thought of as a generalized correlation analogous to the linear correlation coefficient, but sensitive to any relationship, not just linear dependence [13,14]. But it can be seen from the definition of the mutual information that it does not ranges in a definite closed interval as the correlation coefficient does, which ranges in $[0, 1]$ with 0 indicates the minimum linear correlation and 1 indicates the maximum. In this section, we will develop a revised version of the mutual information, which will be sensitive to the general correlation of two variables as the mutual information does, while ranges from a closed interval $[0, 1]$ as the correlation coefficient does.

Considering two discrete variables $X = \{x_i\}_{1 \le i \le N}$ and $Y = \{y_i\}_{1 \le i \le N}$, they are first resorted in ascending order and placed into b ranks with first N/b samples in the first rank, the second N/b samples in the second rank, and so on. Second, the sample pairs, $\{(x_i, y_i)\}_{1 \le i \le N}$, are placed into a $b \times b$ rank grids by comparing the sample pairs to the rank sequences of X and Y.

The revised joint entropy of the two variables X and Y is defined as

$$H^{\mathrm{r}}(X, Y) = -\sum_{i=1}^{b}\sum_{j=1}^{b} \frac{n_{ij}}{N} \log_b \frac{n_{ij}}{N}, \tag{4}$$

where n_{ij} is the number of samples distributed in the ijth rank grid. And the nonlinear correlation coefficient is defined as

$$\mathrm{NCC}(X;Y) = H^{\mathrm{r}}(X) + H^{\mathrm{r}}(Y) - H^{\mathrm{r}}(X,Y), \tag{5}$$

where $H^{\mathrm{r}}(X)$ is the revised entropy of the variable X, which is defined as

$$H^{\mathrm{r}}(X) = -\sum_{i=1}^{b} \frac{n_i}{N} \log_b \frac{n_i}{N}. \tag{6}$$

Notice that the number of samples distributed into each rank of X and Y is invariant, and the total number of sample pairs is N, so the nonlinear correlation coefficient, i.e. Eq. (5), can be rewritten as

$$\mathrm{NCC}(X,Y) = 2 + \sum_{i=1}^{b^2} \frac{n_{ij}}{N} \log_b \frac{n_{ij}}{N}. \tag{7}$$

The nonlinear correlation coefficient not only is sensitive to the nonlinear correlation of two variables, but also can describe this relationship with a number ranges from the closed interval [0, 1], with 0 indicates the minimum general correlation and 1 indicates the maximum one. In the maximum correlation condition, sample sequences of the two variables are exactly the same, i.e. $x_i = y_i$ ($i = 1, 2, \ldots, N$). So the nonlinear correlation coefficient is $I^{\mathrm{r}}(X;Y) = 2 + \sum_{i=1}^{b^2} p_i \log_b p_i = 2 + b * \frac{N/b}{N} \log_b \frac{N/b}{N} = 1$ Under the minimum correlation situation, while the sample pairs distributed equally into the $b \times b$ ranks. So $I^{\mathrm{r}}(X;Y) = 2 + \sum_{i=1}^{b^2} p_i \log_b p_i = 2 + b^2 * \frac{N/b^2}{N} \log_b \frac{N/b^2}{N} = 0$.

For multivariate situation, the general relation between every two variables can be obtained according to the definition of nonlinear correlation coefficients, thus the nonlinear correlation matrix of the concerned K variables can be written as

$$R^N = \{\mathrm{NCC}_{ij}\}_{1 \le i \le K, 1 \le j \le K}, \tag{8}$$

where NCC_{ij} denotes the nonlinear correlation coefficient of the ith and jth variable. As a variable is completely the same as itself, $\mathrm{NCC}_{ij} = 1$ ($i = j$, $1 \le i \le K$, $1 \le j \le K$).

The diagonal element of R, $r_{i,j} = 1$ ($i = j$, $i \le N$, $j \le N$), represents the autocorrelation of each variable. The rest of the element of R, $0 \le r_{i,j} \le 1$ ($i \ne j$, $i \le N$, $j \le N$), denotes the correlation of ith and jth variable. When the variables have no relation with each other, R is unit matrix. In this case, the multi-variables have the weakest relation. When all the variables have the strongest correlation with each other, each element of R equals to 1. In this situation, the correlation of the multi-variables is also the strongest.

The general relation of the concerned K variables is implied in R^N. In order to quantitatively measure it, the nonlinear joint entropy H_{R^N} is defined as following:

$$H_{R^N} = -\sum_{i=1}^{K} \frac{\lambda_i^{R^N}}{K} \log_K \frac{\lambda_i^{R^N}}{K}, \tag{9}$$

where $\lambda_i^{R^N}$ ($i = 1, 2, \ldots, K$) are the eigenvalues of the nonlinear correlation matrix. According to matrix eigenvalues theory, it can be educed that $0 \le \lambda_i^{R^N} \le K$ ($i = 1, 2, \ldots, K$) and $\sum_{i=1}^{K} \lambda_i^{R^K} = K$.

The nonlinear correlation information entropy I_{R^N}, used as a nonlinear correlation measure of the concerned variables, is defined as

$$I_{R^N} = 1 - H_{R^N} = 1 + \sum_{i=1}^{K} \frac{\lambda_i^{R^N}}{K} \log_K \frac{\lambda_i^{R^N}}{K}. \tag{10}$$

Nonlinear correlation information entropy has some excellent characters (mathematical properties) that further prove its suitableness as a measure for the nonlinear correlation of the multi-variables.

First, it remains unchanged when the positions of the K variables are changed. Because the changes of the position of the variables are just a similar transform to the nonlinear correlation matrix, which will not change the eigenvalues of the matrix. It means that the nonlinear correlation information entropy will not be changed.

Second, it ranges from a closed interval [0, 1], with 0 indicates the minimum nonlinear correlation among the K variables concerned, while 1 indicates the maximum. In the minimum nonlinear situation, the nonlinear correlation coefficient of each two different variables is zero. This leads to the nonlinear correlation matrix become an identity matrix, i.e., $\lambda_i^{R^N} = 1\,(i = 1, 2, \ldots, K)$. As a result, the nonlinear correlation information entropy equals to zero. In the maximum correlation situation, the nonlinear correlation coefficient of each two variables equals to 1. This leads to every element of the nonlinear correlation matrix equals to 1, i.e. the eigenvalues $\lambda_i^{R^N} = 0\,(i = 1, 2, \ldots, K-1)$ and $\lambda_K^{R^N} = K$. In this way, the nonlinear correlation information entropy equals to 1.

Third, it is sensitive to the general relations of the K variables concerned, not merely the linear relations. This characteristic will be testified by the following numerical simulations. Fig. 1 shows the relations of three random distributions, i.e. uniform distribution, normal distribution, exponential distribution, and corresponding NCIE. As the three variables concerned are randomly distributed, their relations is weak, so the NCIE is also very little. Fig. 1

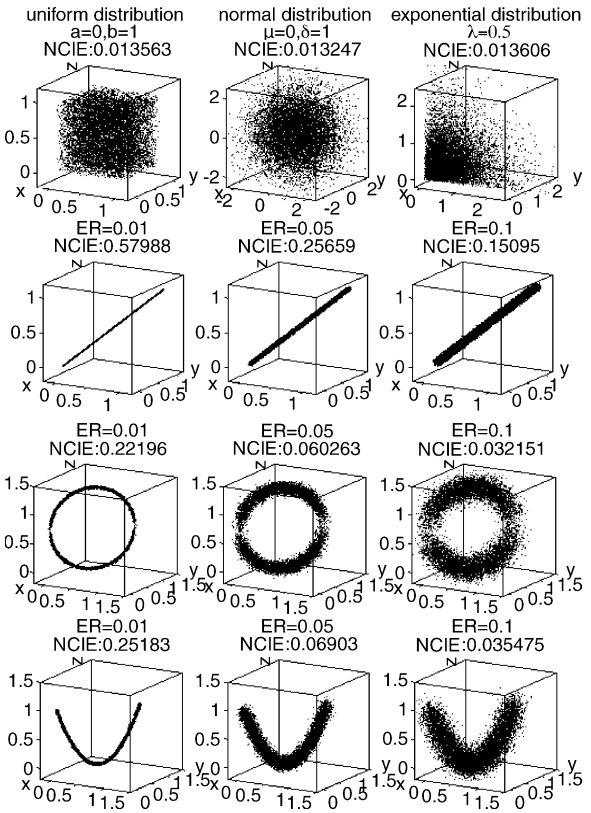

Fig. 1. NCIE of three random variables of three classical distributions, i.e. uniform distribution, normal distribution and exponential distribution, and three common relations, i.e. linear relation, circular relation and square relation with different noises added to the variables. The number of values of each variable $N = 10{,}000$ and the bin number $b = 100$.

also shows the relations of three common functions, i.e. linear, circle and square relations of three variables. In Fig. 1, the noises with different amplitudes are added to the functions in order to generate different correlation degree. The NCIE of the corresponding relation can also be found in the figures.

From the figures, it can be seen that as the amplitude of the added noise increases, the correlation degree of the concerned three variables decreases and their NCIE also decreases. This result conforms to our definition of NCIE, which states that larger NCIE indicates stronger correlation. Meanwhile, the tendency that stronger correlation has larger NCIE can be obviously found.

3. Examples

Multi-variable time series generated from a liner autoregressive (AR) model, and three-dimensional nonlinear chaotic Lorenz system, and a logistic function are used as examples of the numerically generated data of the known origin in order to demonstrate the proposed concepts of nonlinear correlation coefficients and nonlinear correlation information entropy.

3.1. Linear autoregression

A thousand samples of the three-variable series $\{x(t), y(t), z(t)\}$ are generated by the linear AR model:

$$x(t) = 0.9x(t-1) + \sigma_1(t), \qquad y(t) = 0.9x(t-1) + 0.9y(t-1) + \sigma_2(t),$$
$$z(t) = 0.9x(t-1) + 0.9y(t-1) + 0.9z(t-1) + \sigma_3(t), \tag{11}$$

where $\sigma_1(t)$, $\sigma_2(t)$ and $\sigma_3(t)$ are Gaussian deviates with zero mean and unit variance. Fig. 2 displays the samples of three variables. Fig. 3 displays the nonlinear correlation coefficients and mutual information between every two variables. In the figures we can find that the stronger relationship leads to the larger mutual information and also results in the larger nonlinear correlation coefficient. Moreover, the trends of NCC coincide with that of the mutual

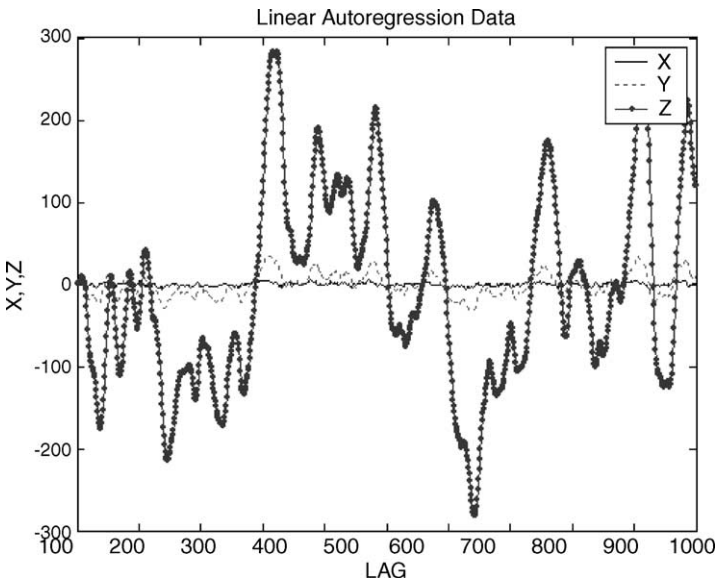

Fig. 2. Samples of three variables of the linear autoregressive model.

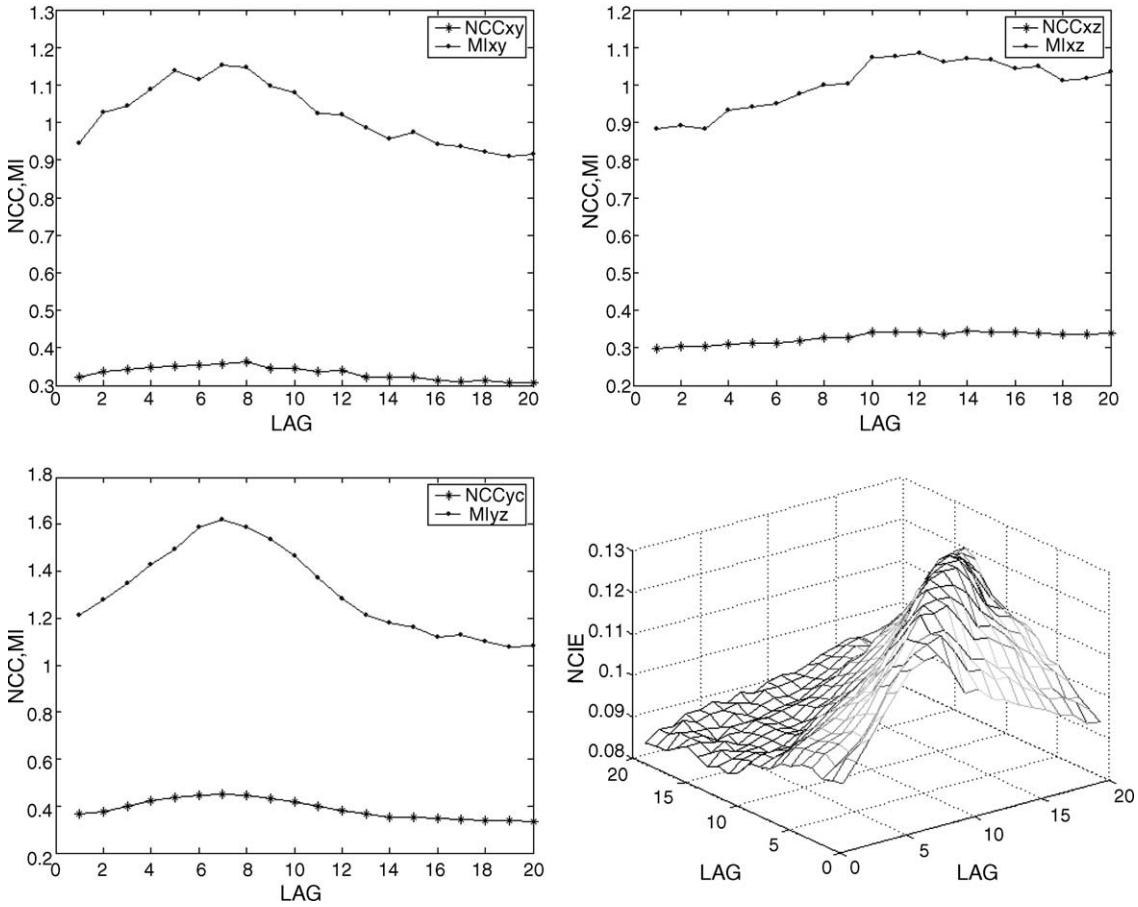

Fig. 3. Nonlinear correlation coefficients and mutual information between every two variables with latter of the variables lags from 1 to 20 steps, and NCIE of the three variables with variable Y and Z lag from 1 to 20 steps.

information. And in the 3D NCIE figure, we can easily find the lag steps, at which the three variables related with each other the most strongly.

3.2. Lorenz system

A thousand samples of the three-variable series $\{x(t), y(t), z(t)\}$ are obtained from the chaotic Lorenz system [15]:

$$\left(\frac{\mathrm{d}x}{\mathrm{d}t}, \frac{\mathrm{d}y}{\mathrm{d}t}, \frac{\mathrm{d}z}{\mathrm{d}t}\right) = \left(10(y-x), 28x - y - xz, \frac{xy - 8z}{3}\right) \tag{12}$$

with the initial values (15.34, 13.68, 37.91). Fig. 4 shows the three-variable samples in three-dimensional figure.

Table 1 lists the nonlinear correlation coefficients and mutual information of each two variables for comparison. It can be obviously found that the variables X and Y have the maximum relationship, as the NCC_{xy} and MI_{xy} are the largest among the three results. However the NCC_{xy} value of 0.43295 indicates a correlation degree, which can be inferred from the NCC value of 1 indicates two variables are completely correlated with each other, and NCC value of 0 indicates two variables have no relation with each other, while MI_{xy} has no such indications as it does not

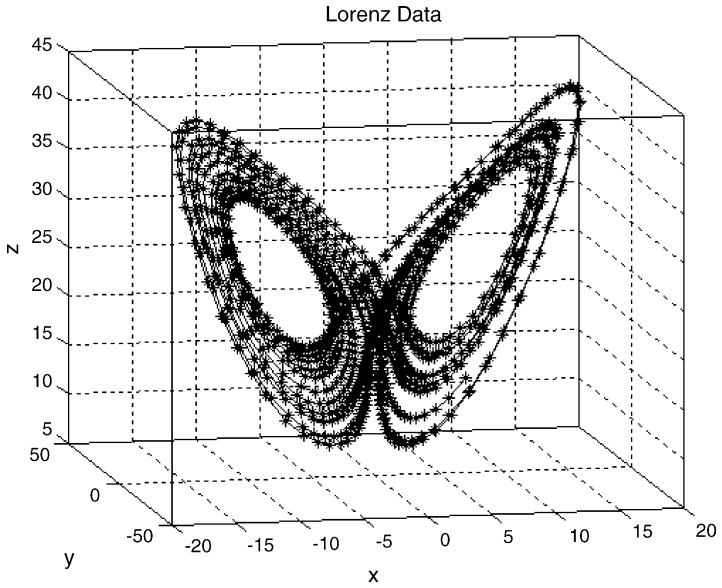

Fig. 4. Three-variable samples of the chaotic Lorenz system in 3-dimension.

range in a closed interval. Also, the nonlinear correlation information entropy of the three-variable NCIE = 0.10753 is listed in the table, which indicates a moderate correlation.

3.3. Numerical simulation

In order to testify the effectiveness of the proposed nonlinear correlation information entropy used to measure the nonlinear correlation among multi-variables, we conduct a numerical simulation using the data generated by a well-known logistic equation, which is thoroughly studied and applied in [13]. The data generation equation used is

$$x_t = 4x_{t-4}(1 - x_{t-4}). \tag{13}$$

The reason for the use of this logistic equation is that the distinct peaks in the lagged self-mutual information can be obtained. The logistic equation produces a chaotic time series and each series in the simulation has 1000 values. Table 2 shows the nonlinear correlation coefficients between time series lagged by i and j ($1 \leq j, j \leq 10$) steps from the original (we call the series lagged by i steps from original band i).

Table 1
Nonlinear correlation coefficients and mutual information of every two variables of the Lorenz system

NCC_{xy}	0.43295
NCC_{xz}	0.32456
NCC_{yz}	0.28844
MI_{xy}	1.43841
MI_{xz}	1.17631
MI_{yz}	0.86114
NCIE	0.10753

Table 2
Nonlinear correlation coefficients between time series lagged by i and j ($1 \leq i, j \leq 10$) steps from the original

Band	1	2	3	4	5	6	7	8	9	10
1	1	0.51363	0.51290	0.51108	0.81729	0.50745	0.51425	0.50554	0.73123	0.51065
2	–	1	0.51394	0.51457	0.51342	0.81371	0.50788	0.51458	0.50773	0.73251
3	–	–	1	0.51358	0.51299	0.51218	0.81645	0.50927	0.51495	0.50688
4	–	–	–	1	0.51475	0.51565	0.51260	0.81341	0.50832	0.51482
5	–	–	–	–	1	0.51656	0.51481	0.51209	0.81283	0.50828
6	–	–	–	–	–	1	0.51609	0.51508	0.51303	0.81276
7	–	–	–	–	–	–	1	0.51559	0.51433	0.51288
8	–	–	–	–	–	–	–	1	0.51761	0.51437
9	–	–	–	–	–	–	–	–	1	0.51807
10	–	–	–	–	–	–	–	–	–	1

Table 3
Nonlinear correlation information entropy of some typical combinations of bands

Group ID	Series combination	NCIE
1	1, 2, 3, 4	0.23732
2	1, 2, 3, 5	0.31285
3	1, 2, 3, 4, 9	0.27880
4	1, 2, 3, 4, 5	0.29850
5	1, 2, 3, 5, 9	0.36321

The simulation of nonlinear correlation information entropy is conducted on the generated time series. The generated series has maximum correlation with the version of itself lagged by four steps. The nonlinear correlation information entropy (NCIE) of some typical combinations of time series is shown in Table 3.

According to Tables 2 and 3, series 1, 2, 3 and 4 in group 1 correlate with each other ordinarily (comparing to series 1 and 5, or series 2 and 6, etc.), and the NCIE of this group is comparatively little, which also implies this group of series correlate less with each other. Group 2 contains a pair of most correlated series 1 and 5, this makes the correlation degree of this group larger than that of group 1. And NCIE of this group is also larger than the first group, which also implies more correlation in this group. Comparing groups 3, 4, and 5, it can be found that the correlation degree of the three groups increases, and the value of NCIE of the groups also increases. So, it can be concluded that, if the series in a group have more correlation, the NCIE of this group is larger than others. This conclusion completely complies with our definition of CIE.

4. Conclusion

It has been shown that the proposed nonlinear correlation coefficient and nonlinear correlation information entropy can be used to measure the general relationship between two time sequences and among a multi-variable data set with a number from the closed interval $[0, 1]$, with the value 0 indicates the minimum relationship and value 1 indicates the maximum. The two concepts can be considered as the extensions to the classical correlation coefficient and mutual information. But extensive study of dynamical systems using ranks has revealed that the conversion of the data still has some limitations such as it cannot well represent ‘rare and/or large events’. For example, an occasional large excursion into the third dimension may exist in the Rossler system and it is only weakly captured. So, the correlation between the original trajectory and its two-dimensional projection is easily overestimated.

Acknowledgements

We thank the reviewers for their comments and suggestions. This study is supported by National Natural Science Foundation of China (Project 60272073) and by the Scientific Research Foundation of Harbin Institute of Technology (Project HIT.2002.11).

References

[1] R.S. Shaw, Strange attractors, chaotic behavior, and information flow, Z. Naturforch. 36a (1981) 80–112.
[2] J.-P. Eckmann, D. Ruelle, Ergodic theory of chaos and strange attractors, Rev. Mod. Phys. 57 (1985) 617–656.
[3] C. Shannon, W. Waver, The Mathematical Theory of Communication, University of Illinois Press, Chicago, IL, 1949.
[4] P. Grassberger, I. Procaccia, Estimation of the Kolmogorov entropy from a chaotic signal, Phys. Rev. A 28 (1983) 2591–2593.
[5] A. Cohen, I. Procaccia, Computing the Kolmogorov entropy from time signals of dissipative and conservative dynamical systems, Phys. Rev. A 31 (1985) 1872–1882.
[6] K. Pawelzik, H.G. Schuster, Generalized dimensions and entropies from a measured time series, Phys. Rev. A 35 (1987) 481–484.
[7] A. Renyi, Probability Theory, North-Holland, Amsterdam, 1970.
[8] A.M. Fraser, H.L. Swinney, Independent coordinates for strange attractors from mutual information, Phys. Rev. A 33 (1986) 1134–1140.
[9] A.M. Fraser, Information and entropy in strange attractors, IEEE Trans. Inform. Theory IT-35 (1989) 245–262.
[10] M. Palus, V. Albrecht, I. Dvorak, Information theoretic test for nonlinearity in time series, Phys. Lett. A 175 (1993) 203–209.
[11] J.D. Farmer, Information dimension and the probabilistic structure of chaos, Z. Naturforch. 37a (1982) 1304–1325.
[12] D. Prichard, J. Theiler, Generalized redundancies for time series analysis, Physica D 84 (1995) 476–493.
[13] M.S. Roulston, Significance testing of information theoretic functionals, Physica D 110 (1997) 62–66.
[14] M.S. Roulston, Estimating the errors on measured entropy and mutual information, Physica D 125 (1999) 285–294.
[15] M. Palus, Detecting nonlinearity in multivariate time series, Phys. Lett. A 213 (1996) 138–147.