The 7th International Symposium on Operations Research and Its Applications (ISORA'08)
Lijiang, China, October 31–Novemver 3, 2008

Improved Model and Algorithm for Determining Correlation of Trend Sequences

Yin Chen[1, *]

[1]School of Science, South China Agricultural University, Guangzhou 510642, China

Abstract The traditional method to construct the trend sequence and calculate the correlation index is introduced. Then disadvantage of the traditional method is briefly described. The model and improved algorithm are proposed to settle the disadvantage. For the reciprocity of two systems, the information concealing in the series of dynamic data can be obtained by using this model and improved algorithm.

Keywords Trend Sequence; Generalized Correlation Index; Threshold Value; Dynamic Data

1 Introduction

As a matter of fact, a system is formed by the reciprocity between its elements. Then the system has mutual effect with other systems. Therefore, for many systems, the change of its dynamic data is not only owing to itself, but also because of dynamic data of other systems.

So the evolution of a system is influenced by two parts. One is the reciprocity inside. The other is the mutual effect with other systems. The mutual effect between systems is achieved by getting help from the circulating of matter, energy and information. The circle process is as follows. Firstly, matter, energy and information from other systems are input into the system to participate in the reciprocity inside. Secondly, the system output new matter, energy and information to external. On the other hand, these matter, energy and information may be input into other systems to influence the evolution of other systems.[1]

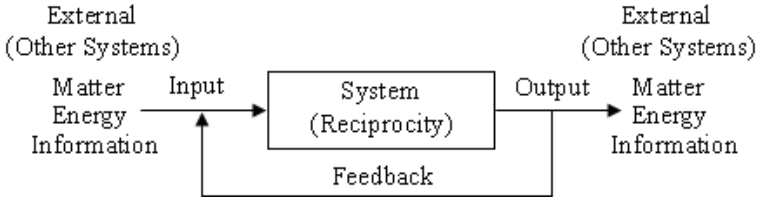

Figure 1: Process of mutual effect between systems

*Email: gdstchenyin@126.com

By observing the phenomena of nature, we can get series of dynamic data from different natural systems. These data, which change along with time, represent the evolution of the systems. In order to know whether there is interrelation between two different systems or not, people process the series of dynamic data with diverse methods. The model and algorithm in this paper also can find out the correlation hiding in the dynamic data effectively.

2 Theories of Trend Sequence[2]

Trend sequence is an abstract of series of dynamic data. Instead of numerical value, there are symbols in trend sequence. So trend sequence is more intuitional and conciseness than the series of dynamic data. The definition about trend sequence is as follows.

Definition 2.1 Trend Sequence

Set a series of dynamic data $X = \{x(n) | x(n) \in R, 1 \leq n \leq N\} \in R^N$, where N denotes its length. Set a discrete limited set $D = \{d_k | 1 \leq k \leq M\}$, D is set of trend symbol, where d_k is trend symbol. Set a mapping $f : R^N \to D^N$, f is a trend function. Given trend symbol set D and trend function f, the corresponding trend sequence of X is $T = f(X) \in D^N$.

The general way to convert series of dynamic data into trend sequence is based on its first-order difference. The specific steps are as follows. For a series of dynamic data X, firstly, compute its first-order difference.

$$\Delta x(n) = x(n+1) - x(n), \quad n = 1, 2, \cdots, N,$$

where N is the length of X.

Then set u,s and d as trend symbols. According to the value of $\Delta x(n)$, define trend function as follows.

$$t(n) = \begin{cases} u, & \Delta x(n) \geq 0 \\ s, & \Delta x(n) = 0 \\ d, & \Delta x(n) < 0 \end{cases} \tag{1}$$

Then the series of dynamic data X is converted into trend sequence T. Graphically, u shows ascending. s shows steady-going. And d shows decline.

Furthermore, the definition of trend correlation index, which can be used to describe and quantify the relationship of two trend sequences, is as follows.

Definition 2.2 Trend Correlation Index

Given two series of dynamic data whose lengths are equal, convert these two series into trend sequences with the same trend function. Then match these two trend sequences. And the ratio between the amount of consistent trend symbols and the length of the trend sequence is trend correlation index of these two trend sequences.

The method to calculate the correlation index of trend sequences is as follows.

Given trend sequences $T_1, T_2 \in D^N$, the correlation coefficient of T_1 and T_2 is

$$S'(T_1, T_2) = \sum_{n=1}^{N} \delta(T_1(n), T_2(n)),$$

where $\delta(a,b) = \begin{cases} 1, & a = b \\ 0, & a \neq b \end{cases}$.

So the trend correlation index is

$$S(T_1,T_2) = \frac{S'(T_1,T_2)}{N}.$$

On the one hand, if the trend correlation index of T_1 and T_2 is close to 1, it means T_1 and T_2 have the same tendency most of time. Then these two sequences are correlative and the evolutions of them are similar. On the other hand, if the trend correlation index of T_1 and T_2 is close to 0, it means T_1 and T_2 having opposite tendency most of time. Then these two sequences are correlative and the evolutions of them are contrary.

Set the threshold value of similar correlation is$\alpha \in [0,1]$, and the threshold valve of opposite correlation is$\beta \in [0,1]$. And they meet $\alpha > \beta$.

If there is $S(T_1,T_2) \geq \alpha$, it means the correlation of T_1 and T_2 is remarkable, and they are similarly correlative. So the series of dynamic data X_1 and X_2 evolve similarly.

If there is $S(T_1,T_2) \leq \beta$, it means the correlation of T_1 and T_2 is remarkable, and they are oppositely correlative. So the series of dynamic data X_1 and X_2 evolve oppositely.

If there is $\beta < S(T_1,T_2) < \alpha$, it means the correlation of T_1 and T_2 is unremarkable. So the evolution of X_1 and X_2 are unrelated.

3 Disadvantage of Traditional Method

By using the above method, we can know whether there is interrelation between any two different sequences or not with their trend correlation index. This model is simple, and there is only a small quantity of calculation in it. However, it also has disadvantage, which interfere its accuracy. Specifically, for two trend sequences, if the stagger arrangement of one trend sequence correlates with the other one, their correlation cannot be find out by using the traditional method.

For instance, two series of dynamic data are as follows.

$$\begin{cases} X_1(n) = \sin\frac{n\pi}{4} \\ X_2(n) = \sin\frac{(n-2)\pi}{4} \end{cases}$$

Graph of X_1 and X_2 can be get according to the above formulas.

Obviously, X_1 and X_2 are correlative. Because when the curve of X_1 parallel moving two units of axis to the right, the curves of X_1 and X_2 are overlapping.

Now the traditional method is used to calculate the correlation index. Firstly,X_1 and X_2 are changed into trend sequences T_1 and T_2 by computing the first-order difference and using the formula (1). Then the trend correlation index of T_1 and T_2 is obtained by using the traditional algorithm. And set $\alpha = 0.8, \beta = 0.2$.

By observing Figure 2, the number of times that T_1 and T_2 having the same trend symbol is equal to the number of times that T_1 and T_2 having different trend symbol. So the trend correlation index of these two trend sequences is $S(T_1,T_2) = 0.5$. So there is $\beta < S(T_1,T_2) < \alpha$. It means the correlation of T_1 and T_2 is unremarkable. Therefore, the evolution of X_1 and X_2 are unrelated. Evidently, the correlation of T_1 and T_2 cannot be find out in this way.

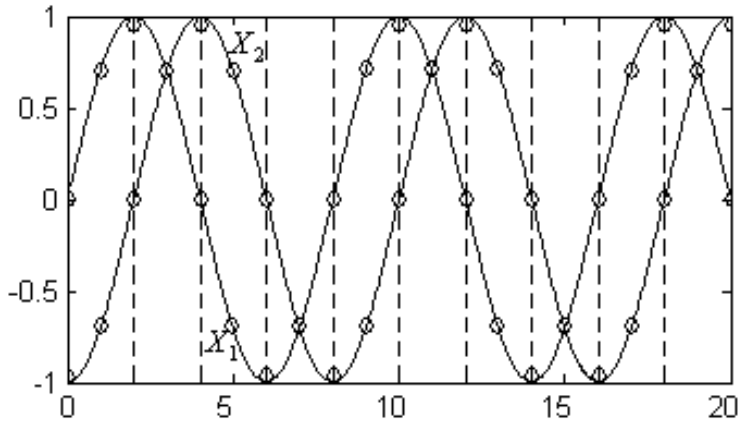

Figure 2: Curve sketching of X_1 and X_2

4 The Improved Model and Algorithm

According to the disadvantage of traditional method which are mentioned above, the improved model and algorithm are proposed. They can settle the disadvantage effectively.

Briefly, the main idea of the improved method is not only to count the trend correlation index of two original trend sequences, but also to further calculate the trend correlation index after transposition according to time unit. By this way, more information can be excavated than before. Therefore, this improved method can describe the relationship of the trend sequences better.

In order to make the improved model and algorithm clearly understood, a new concept, generalized trend correlation index, is proposed. Essentially, it is a modified form of trend correlation index.

Definition 4.1 Generalized Trend Correlation Index

For two series of dynamic data that having equal lengths, convert these two series into trend sequences with the same trend function. Then transpose one series unit by unit while fixing the other one. Match two series when a series transposes a unit, and calculate their trend correlation index. The maximum of all trend correlation indexes which are gained during transposition is the correlation index of these two trend sequences. This maximum is called generalized trend correlation index.

Specifically, the improved method to calculate the generalized correlation index of trend sequences is as follows.

Given two trend sequences, they both consist of N terms. When one trend sequence transposes i units relatively to the other sequence, the correlation index of these two sequences is

$$S_i(T_1, T_2) = \frac{S'_i(T_1, T_2)}{N}, \tag{2}$$

where $S'_i(T_1, T_2) = \sum_{n=i}^{N} \delta(T_1(n), T_2(n))$.

In the original method, it is mentioned previously that the series are similarly correlative if their correlation index is more than the threshold value of similar correlation, and they are oppositely correlative if their correlation index is less than the threshold value of opposite correlation.

Because it is asked to count all correlation indexes after transposition in the improved method, the calculation becomes fairly complex. In order to simplify the calculation, further modify the design procedure, which is based on formula (2), to compute the correlation index.

Define that

$$S_i = \max\{S_i(T_1,T_2), 1-S_i(T_1,T_2)\}. \tag{3}$$

Here S_i is the correlation index of these two sequences when one trend sequence transposes i units relatively to the other sequence,

Correspondingly, amalgamate the threshold value of similar correlation and that of opposite correlation to be a new measuring value, which is called threshold value of correlation. The concept of threshold value of correlation is proposed according to generalized trend correlation index. It is used to judge whether the correlation between trend sequences is remarkable or not.

Definition 4.2 Threshold Value of Correlation Index

Given a value λ, if it is considered that the correlation of two trend sequences is remarkable when the correlation index is larger than λ or equal to λ, and the correlation is unremarkable when the correlation index is smaller than λ, then this λ is the threshold value of correlation index.

In addition, the maximum of units for one trend sequence to transpose is t.

$$t = [(1-\lambda)N]$$

Namely, in formula (2) and (3), $i = 0,1,2,\cdots,t$. Because when $i > t$, a trend sequence moves more than $(1-\lambda)N$ units. Correspondingly, the number of the rest symbols to match the other sequence is less than λN. In such a case, even if the rest symbols of two sequence to match are all identical or all opposite, the value of $S_i'(T_1,T_2)$ is less than λN. So $S_i(T_1,T_2) = \frac{S_i'(T_1,T_2)}{N} < \lambda$. Therefore, when $i > t$, the correlation index of two trend sequences is less than the threshold value, which means the maximum of units for one trend sequence to transpose is t.

Hence, the model for calculating the generalized correlation index of two trend sequences is

$$S = \max_{0\le i\le t}\{S_i\}.$$

If $S \ge \lambda$, the correlation of two trend sequences is remarkable. If $S < \lambda$, the correlation of two trend sequences is unremarkable.

Steps of the improved algorithm are as follows.

Step 1.Define the set of trend symbol D and the corresponding function f. Then change the series of dynamic data X_1, X_2 into trend sequences T_1, T_2.

Step 2.Set $i=0$. And set the numerical value of λ.

Step 3.Calculate $t=[(1-\lambda)N]$.

Step 4.Fix one trend sequence. And transpose the other trend sequence i units forward relatively to the fixed sequence.

Step 5.Calculate $S_i'(T_1,T_2)=\sum_{n=i}^{N}\delta(T_1(n),T_2(n))$.

Step 6.Calculate $S_i(T_1,T_2)=\frac{S_i'(T_1,T_2)}{N}$.

Step 7.Calculate $S_i=\max\{S_i(T_1,T_2),1-S_i(T_1,T_2)\}$.

Step 8.If $i<t$, $i=i+1$. If not, turn to *Step* 4.

Step 9.Calculate $S=\max_{0\le i\le t}\{S_i\}$.

Step 10.If $S\ge\lambda$, the correlation of two trend sequences is remarkable. If $S<\lambda$, the correlation of two trend sequences is unremarkable.

Review the foregoing example. Set $\lambda=0.8$. So if there is no transposition, $S_0(T_1,T_2)=0.5<\lambda$. However, When parallel transpose T_1 two units forward, T_1 is the same with T_2. And the generalized correlation index in this moment is $S_2(T_1,T_2)=1$, which is maximum. So $S=1>\lambda$. It means these two trend sequence are correlative after transposition. So the correlation hiding in the series of dynamic data is found out by this way.

5 Concluding Remarks

For the reciprocity of two systems, the information concealing in the series of dynamic data can be obtained by using this improved model and algorithm. Especially for the trend sequences whose stagger arrangements are correlative, this improved method can find out the relation between variables.

The situation that one trend sequence is correlative with the stagger arrangement of the other trend sequence can be understood as that there is time difference in the reciprocity of systems. It means the influence to the system that X_2 representing, which is caused by the output of the system that X_1 representing, does not have the immediate effect. It acts after a period of time. Briefly, it shows that the output of one system influences another system after a period of time. Therefore, this model and the improved algorithm compensate for disadvantage of the traditional algorithm. It is more useful and effective than the traditional one.

References

[1] Guangrong Zeng, Discussing the Basic Form and Course of Interaction of System, *System Dialectic Transaction*, 41-45, Vol.12, No.1, Jan. 2004, (In Chinese).

[2] Siyu Guo, Research of data mining on dynamic data, *ZheJiang University Ph.D. Dissertation*, 25, 37, 2002, (In Chinese).