Acta Technica 62 No. 2B/2017, 349–360 

# Analysis algorithm for internet of things big data based on multi-granularity functional

SUN JING[2]

**Abstract.** The Internet of things big data is a kind of important object of data mining. In Internet of things big data analysis process, if the time difference of the data is not considered, the misjudgment of the correlation will occur. First of all, from the perspective of the time warping, the reason and characteristics of two types of correlation errors were analyzed; Then, according to the asymptotic distribution of correlation coefficients, the boundary of correlation coefficients in a certain degree of significance level was obtained. The correlation method based on time shift sequence correlation coefficient characteristic was obtained by integrating both of them; finally, the multi-granularity functional model based on the maximum correlation coefficient was put forward, which has a wider application scope than AISE principle. The model is based on Multi-granularity maximize functional (MGMF) algorithm for solving the time warping function. The numerical experiment results in constructing data and real data show that, in the spurious regression identification, the correlation method is more effective than 3 kinds of conventional correlation coefficients and Granger causality test; in most cases the proposed MGMF algorithm is better than continuous monotone registration method (CMRM), self-modeling registration (SMR) and maximum likelihood registration (MLR).

**Key words.** Spurious regression, time warping; correlation, curve registration, Internet of things big data.

## 1. Introduction

The Internet of things big data is one of the most common data type of in data mining, which has been applied in many fields, such as the monthly runoff of certain ricer, the average monthly temperature and precipitation of local area, China's consumer price index (CPI) and gross domestic product (GDP), the seismic sequences of a number of observation points when the earthquake happened and so

[1]This work was support by the Xi'an Innovation Special Fund of science and technology project, CXY1352WL13.

[2]School of Mechanical and Materials Engineering, Xi'an University, Xi'an, Shaanxi, 710065, China

on. The analysis of these Internet of things big data can get some useful conclusions. For example: through the study of river history flow, temperature and precipitation characteristics, the level of flood forecast can be effectively improved; the utilization of CPI and GDP can analyze the extent of inflation and economic development momentum in countries or regions; according to the seismic wave sequences, the source and magnitude of earthquake can be accurately located [1]. Certain non-Internet of things big data can also be transformed by Internet of things big data for analysis, for example: the distance from the edge to the centroid of leaf can be used to describe its characteristics, and a series of data from different angles can be obtained, and then the type of leaf can be identified [2].

In the Internet of things big data analysis process, if the time difference of the data was not considered, it is easy to be affected by intuition or prejudice, and introduce erroneous judgement of the correlation; but the time difference does not consider the correlation sequence is of no significance. That is to say, determining the sequence correlation needs to consider the time difference. It is necessary to consider the time difference and correlation data, so the correlation and the time sequence among data are restricted with each other. At present, the correlation analysis of Internet of things big data is faced with some problems, such as the data relation is complex, data contains noise, missing data or abnormal data [3]. Homogeneous data (data with the same source or attribute, such as the seismic wave data with same earthquake obtained from multi areas) have natural similarity, which do not need to determine the correlation and does not exist correlation and time difference constraint problem. It is generally used for classification or clustering. While for the heterogeneous data (data with the different sources or attributes, such as precipitation and river runoff, CPI and GDP), it is necessary to determine their relevance. If there does exist relevance, regression analysis can be carried out, etc. Therefore, the main object of the Internet of things big data analysis is heterogeneous data.

The main reason of correlation error happened is the two groups of Internet of things big data have time warping. The coordination of both can be realized as long as time conversion is conducted. In reality, it is generally nonlinear time warping or dynamic time warping (DTW) [4, 5]. This requires the help of functional data analysis (FDA) method to transfer the Internet of things big data into function data for time correction, which is usually called curve registration or curve alignment. The main reason of correlation error happened is the two groups of Internet of things big data have time warping. The coordination of both can be realized as long as time conversion is conducted. In reality, it is generally nonlinear time warping or dynamic time warping (DTW) [4, 5]. This requires the help of functional data analysis (FDA) method to transfer the Internet of things big data into function data for time correction, which is usually called curve registration or curve alignment [6]. Kneip and Gasser regarded the extremum as landmark registration. But it is not suitable for the curves with inconspicuous feature points, and the selection of the feature points has great influence on the results. A more general method is: to determine an objective function or a curve, and to align the local features of other curves or minimize some metrics (such as the mean square distance between each curve and the target curve) [7, 8]. Ramsay et al. proposed the continuous monotone registration method (CMRM), to ensure the continuity and consistency of the time warping function [9]. Wang and Gasser put forward a curve registration method based on dynamic time warping model [10, 11]. Kneip et al. used the local nonlinear regulating method about the time warping function to align adjust curves.

The homogeneous data has natural similarity, which can be curve registration function directly. To solve the restriction problem of heterogeneous Internet of things big data correlation and the time difference, the time difference is fixed to judge the correlation of each time shift sequence. Based on sequence correlation, The time difference function is then refined by the curve registration function. When the heterogeneous data are doing correlation judgment, on the one hand, since the sample correlation coefficient has deviation with the overall correlation coefficient, so upper and lower bounds on the overall correlation coefficient are studied; On the other hand, in order to prevent the emergence of two types of correlation errors, starting from the main reason of its occurrence, the characteristics of the two kinds of correlation errors are studied and a method to determine the corresponding correlation judgment method is put forward. The curve registration method suitable for the heterogeneous data is also applicable to homogeneous data, while the criterion suitable for homogeneous data (such as AISE) is not applicable to heterogeneous data (non-uniform dimension and negative correlation etc.). Therefore, mainly according to the characteristics of heterogeneous data, this paper proposes curve registration criterion based on the maximization of correlation coefficient (absolute value), and uses MGMF algorithm to solve the problem.

## 2. Correlation analysis method of curve registration

Because we can only get the sample data in solving practical problems, and it will produce deviation when using samples to estimate the population, this paper uses the sample correlation coefficient to infer the overall correlation coefficient in a certain degree of the boundary; At the same time, in order to prevent the occurrence of two related errors, this paper studies the characteristics of two kinds of shift sequence correlation coefficients, and then excludes two types of correlation errors. These two aspects can be determined to obtain two groups of correlation decision method of the Internet of things big data.

### *2.1. Correlation decision of correlation sequences with time warping*

In order to determine the correlation of the sequence, it is necessary to infer the upper and lower bounds of the overall correlation coefficient. The overall correlation coefficient has upper and lower bounds in a certain degree. It was obtained from two asymptotic distributions of sample correlation coefficient. And then with the combination of the first type of correlation error, the correlation decision method of correlation sequences with time warping was obtained.

*2.1.1. Boundary of the correlation coefficient.* Pearson correlation coefficient is the most common metric form for measuring the sequence correlation. If there are two groups of corresponding data $\{(x_i, y_i), i = 1, 2, ..., n\}$ ($n$ being the sample capacity), which are the samples come from binary normal population $(x, y) \sim N(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$, then the sample correlation coefficient is

$$\hat{\rho}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} =$$

$$= \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \cdot \sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} . \tag{1}$$

Here, $\overline{x}$ and $\overline{y}$ are the sample averages of $X$ and $Y$, respectively.

The sample correlation coefficient $\hat{\rho}(X, Y)$ can be used as unbiased and consistent estimator of the correlation coefficient $\rho$ of two normal populations $(X, Y)$. But the correlation coefficient has an obvious disadvantage that its degree closing to 1 is related to the number of data group $n$. It is easy to give an illusion. When $n$ is small, the fluctuation of correlation coefficient is larger. The absolute value of some samples‘ correlation coefficient is close to 1, especially when $n$=2, the absolute value of the correlation coefficient is 1; When $n$ is larger, the absolute value of the correlation coefficient is small. The distribution of results of the sample correlation coefficient, the sample size and the bivariate normal population correlation coefficient were given by many scholars.

Under the assumption that $(X, Y)$ is the bivariate normal population and $\rho$=0, the distribution is as follows:

$$T = \frac{\sqrt{n-2}\hat{\rho}}{\sqrt{1-\hat{\rho}^2}} \sim t(n-2) . \tag{2}$$

When $\rho = \rho_0$, Fisher proposed a more complex probability density function of$\hat{\rho}$. The following asymptotic distribution is obtained after proper transformation:

$$z = \frac{\phi(\hat{\rho}) - \phi(\rho)^{n\to\infty}}{2\sqrt{n-3}} \sim N(0, 1) . \tag{3}$$

Here, $\phi(x) = \ln\frac{1+x}{1-x}$. When the sample size is large, the correlation coefficient can be used to estimate the population correlation coefficient.

Then $n$ samples are extracted in the two normal states, and the following asymptotic distributions are obtained:

$$\sqrt{n}(\hat{\rho} - \rho)^{n\to\infty} \sim N\left(0, \left(1-\rho^2\right)^2\right), \text{ namely } \frac{\sqrt{n}(\hat{\rho} - \rho)^{n\to\infty}}{(1-\rho^2)} \sim N(0, 1) . \tag{4}$$

In this paper, the population correlation coefficients are estimated based on the two asymptotic distributions.

Since $\phi(x)$ is a monotonically increasing function in Formula (2), then

- When $\rho \geq \hat{\rho}$, then

$$P\left\{\rho \leq \phi^{-1}\left[\phi\left(\hat{\rho}\right)+2 z_{1 \frac{a}{2}} \cdot \sqrt{n-3}\right]\right\}=1-a\,. \tag{5}$$

- When $\rho \leq \hat{\rho}$, then

$$P\left\{\rho \geq \phi^{-1}\left[\phi\left(\hat{\rho}\right)-2 z_{1 \frac{a}{2}} \cdot \sqrt{n-3}\right]\right\}=1-a\,. \tag{6}$$

Herein, $\phi^{-1}(x)=\frac{e^{x}-1}{e^{x}+1}$, $Z_a$ is the $a$th quantile of standard normal distribution, namely $P\left(x \leq Z_a\right)$ and random variable $x \sim N(0,1)$.

On the basis of Formula (3), the bounds of the total correlation coefficient are deduced, namely:

- When $\rho \geq \hat{\rho}$, then

$$P\left\{\frac{-\sqrt{n}-Q}{2 z_{1 \frac{a}{2}}} \leq \rho \leq \frac{-\sqrt{n}+Q}{2 z_{1 \frac{a}{2}}}\right\}=1-a\,. \tag{7}$$

- When $\rho \leq \hat{\rho}$, then

$$P\left\{\frac{\sqrt{n}-R}{2 z_{1 \frac{a}{2}}} \leq \rho \leq \frac{\sqrt{n}+R}{2 z_{1 \frac{a}{2}}}\right\}=1-a\,. \tag{8}$$

In (7) and (8)

$$Q=\sqrt{n+4 \sqrt{n} z_{1 \frac{a}{2}} \cdot \hat{\rho}+4 z_{1 \frac{a}{2}}^{2}}$$

and

$$R=\sqrt{n-4 \sqrt{n} z_{1 \frac{a}{2}} \cdot \hat{\rho}+4 z_{1 \frac{a}{2}}^{2}}\,.$$

Synthesizing Formulas (4)–(7), when $a=0.05$, it has the following approximation:

- When $\rho \geq \hat{\rho}$, then

$$\inf_{a=0.05} \rho=\max \left\{\frac{-\sqrt{n}-\sqrt{n+8 \sqrt{n} \cdot \hat{\rho}+16}}{4}, \hat{\rho},-1\right\} \tag{9}$$

and

$$\sup_{a=0.05} \rho=\min \left\{\frac{-\sqrt{n}-\sqrt{n+8 \sqrt{n} \cdot \hat{\rho}+16}}{4}, \phi^{-1}\left[\phi\left(\hat{\rho}\right)+4 \sqrt{n-3}\right], 1\right\}\,. \tag{10}$$

• When $\rho \leq \hat{\rho}$, then

$$\inf_{a=0.05} \rho = \max\left\{ \frac{\sqrt{n} - \sqrt{n - 8\sqrt{n} \cdot \hat{\rho} + 16}}{4}, \phi^{-1}\left[\phi\left(\hat{\rho}\right) - 4\sqrt{n-3}\right], -1 \right\} \tag{11}$$

and

$$\sup_{a=0.05} \rho = \min\left\{ \frac{\sqrt{n} + \sqrt{n - 8\sqrt{n} \cdot \hat{\rho} + 16}}{4}, \hat{\rho}, 1 \right\} . \tag{12}$$

Figure 1 shows the upper and lower bounds of the total correlation coefficient in different sample sizes $n$ and sample correlation coefficients. As can be seen from the graph, the upper and lower bound curves presented in this paper have the following characteristics:

• The larger the sample size, the more compact the upper and lower bounds.

• When the sample size is the same, the upper and lower bound curves are central symmetry.

• The greater the absolute value of the correlation coefficient, the more compact of the upper and lower bounds.

The above characteristics can easily be proved by Formulas (8)–(11).

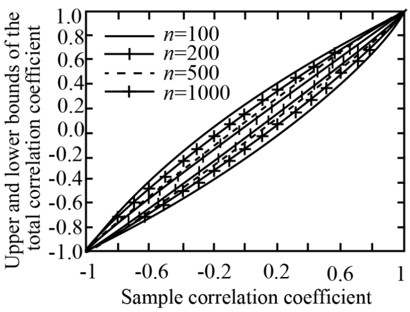


Fig. 1. Bounds of the total correlation coefficient (significance level $\alpha = 0.05$

*2.1.2. Correlation determination method.* In order to describe the characteristics of the sequence, the definition of time-lag series is given.

Assuming there are two sequences $(X, Y) = \{(x_i, y_i), i = 1, 2, ..., n\}$, the following sequence is defined as the time-lag series

$$\begin{aligned} (X_t, Y_{t+m}) &= \{(x_i, y_{i+m}), i = 1, 2, ..., n-m\}, 1 \leq m < n, m \in N^+ , \\ (X_t, Y_{t-m}) &= \{(x_i, y_{i-m}), i = m+1, 2, ..., n\}, 1 \leq m < n, m \in N^+ . \end{aligned} \tag{13}$$

For the first type of regression error, that is, there is no correlation between the related sequences, and just considering the initial sequence, then the inevitable correlation is small; if considering the correlation of time-lag series, then there does exist $m_0 \, (1 \leq |m_0| \ll n, |m_0| \in N^+)$, making the correlation coefficient of $(X_t, Y_{t+m_0})$

larger.

Then the correlation decision of correlation sequences with time warping is obtained: if the correlation coefficient of time-lag series changes with $m$, and reached the maximum at $m_0$, namely $\rho = \hat{\rho}(m)$ in the curve shows obvious convex phenomenon, then the range of correlation coefficient can be estimated according to Formulas (8)–(11). If $|\rho(m_0)| > \rho_0$ (that is more than a given threshold such as 0.6), then the time-lag series $(X_t, Y_{t+m_0})$ has correlation and the curve registration and regression analysis can be carried out.

### *2.2. A curve registration model based on functional correlation coefficient*

The correlation coefficient of time-lag series can determine whether the sequences are correlated or not. If there is correlation in the two sequences, but with time deviation, it needs to align them with curve registration method to eliminate the differences in phase (time axis). For heterogeneous data, when using the AISE criterion, the results will change with the change of dimension. Therefore, we need to put forward a dimensionless criterion to align the heterogeneous data curves.

The Pearson correlation coefficient is a dimensionless measure describing two sequences correlation and similarity, but applies only to describe the correlation of discrete data. The correlation of continuous function can be represented by the inner product. At the same time, in order to make the inner value standardized, the norm of the two functions is divided. The curve registration criterion composed by heterogeneous data can be constructed through the functional correlation coefficient

$$\max_{h(t)} |\rho(x_1^*, x_2)| = \max_{h(t)} \left| \frac{\int_T x_1^*(s)\, x_2(s)\, \mathrm{d}s}{\sqrt{\int_T [x_1^*(s)]^2\, \mathrm{d}s} \cdot \sqrt{\int_T [x_2^*(s)]^2\, \mathrm{d}s}} \right| . \tag{14}$$

Here, $x_1^*(t) = x_1[h(t)]$ means the aligned function. In view of the complexity of the continuous function and the high dimensional feature of the optimization criterion, the corresponding discretization is given in this paper.

Assume the sample sequence of two functional data$x_1(t)$ and $x_2(t)$ at the sampling time points $T = (t_1, t_2, ..., t_n)$ are $x_1(T) = [x_1(t_1), x_1(t_2), ..., x_1(t_n)]$ and $x_2(T) = [x_2(t_1), x_2(t_2), ..., x_2(t_n)]$, and now for function $x_1(t)$ it is necessary to do curve registration as function $x_2(t)$. Order $\Delta = (\delta_1, \delta_2, ..., \delta_n)$ to be the offset of $x_1(t)$ relative to $x_2(t)$ at the time point $T$, namely the time warping function satisfies $h(T) = T + \Delta$, then the aligned time-lag series samples changes into $x_1(T+\Delta) = [x_1(t_1+\delta_1), ..., x_1(t_n+\delta_n)]$. The two groups of functional data sample sequences after aligned should have high correlation. The curve registration problem can be solved by transforming into

$$\max_{\Delta} |\rho[x_1(T+\Delta)], x_2(T)| . \tag{15}$$

Generally, the time warping function has uniform monotonicity, that is, satisfying $t_{i-1} + \delta_{i-1} < t_i + \delta_i < t_{i+1} + \delta_{i+1}$. However, the offset vector will cause the time warping function does not satisfy uniform monotonicity, so $\delta_i^{k+1} \in (bndl_i, bndr_i)$ is limited. Herein, $bndl_i = t_{i-1} + \delta_{i-1}^k - t_i, bndr_i = t_{i+1} + \delta_{i+1}^k - t_i, \delta_i^k$, means the value of $\delta_i$ in the $k$ iteration. In realization, the search range of $\delta_i^{k+1}$ is reduced to closed interval $[bndl_i + p \cdot (bndr - bndl_i), bndr_i - p \cdot (bndr - bndl_i)]$. Herein, $p$ is the constant in (0,0.5) .

In the end, the problem of curve registration is changed to solve the following constrained optimization problem:

$$\begin{cases} \Delta^* = \arg\max_{\Delta} \left|\rho\left[x_1(T+\Delta), x_2(T)\right]\right|, \\ s.t. \delta_i \in [bndl_i + p \cdot (bndr - bndl_i), bndr_i - p \cdot (bndr - bndl_i)] . \end{cases} \tag{16}$$

Finally, the time offset vector $\Delta^*$ is transformed into function form, and the time offset function $d(t)$ is obtained. The corresponding time warping function is $h(t) = d(t) + t$.

## 3. Experimental results and discussion

Based on the simulated data (7 kinds of man-made Internet of things big data and two sets of Sinc function), the proposed correlation decision method and curve registration method were verified. The relevant data for the time warping did the curve registration using this method. On the one hand, the sensitivity of the method to parameters was analyzed; and on the other hand, the existing methods are compared and analyzed.

### 3.1. Correlation decision

*3.1.1. Spurious regression decision.* For the spurious regression decision problem, 7 kinds of main data generating process (DGP) was selected. The correlation coefficient of the time-lag series is used to identify the spurious regression. First of all, the correlations in the 7 groups were analyzed by routine correlation analysis. The results are shown in Table 1.

Model 3 shows that: the correlation coefficients of model 3 and model 5 are relatively small, and have no spurious regression phenomenon. The 3 kinds of correlation coefficients of other groups autocorrelation sequences are relatively high, showing that the conventional correlation coefficient does not have the identify ability to the spurious regression phenomenon of autocorrelation sequence. Model 5 shows that: the Granger causality test can identify the correlation of most of the sequence, but there are still 3 identification errors. The model 7 is two- order autocorrelation sequence, and the bidirectional Granger causality tests were wrong. So for a simple model, Granger causality test can identify spurious regression; when the model is complicated, it can not identify the spurious regression. As the 3 correlation coefficients and Granger causality test can not decide the correlation of autocorrelation sequence, these methods are no longer used in the subsequent experiment for correlation decision.

Figure 2 shows the time-lag series correlation coefficient varying diagram of 7 models. Figure 2 shows that: the absolute values of time-lag series correlation coefficient of model 3 and model 5 are smaller, which will not cause spurious regression. High correlation coefficients are observed in the other 5 models, but the correlation coefficient of time-lag series has little change with $m$. This method can quickly and accurately identify spurious regression. Although this first-order autocorrelation sequence spurious regression discriminant method is given in this paper, it can be seen from the results of model 7: for the two order autocorrelation sequence, it can also obtain reliable correlation results.

Table 1. Comparison of algorithm calculation time (in ms)

| Model | Single integraal order | | Granger causality test (Lag=0) | | | |
|---|---|---|---|---|---|---|
| | $X$ | $Y$ | Cause→Effect | F statistics | Significance | Granger or not Reason |
| 1 | 1 | 1 | $Y \to X$ | 0.006 | 0.993 | No |
| | | | $X \to Y$ | 0.287 | 0.750 | No |
| 2 | 0 | 1 | $Y \to X$ | 1.185 | 0.307 | No |
| | | | $X \to Y$ | 0.019 | 0.980 | No |
| 3 | 1 | 1 | $Y \to X$ | 0.412 | 0.662 | No |
| | | | $X \to Y$ | 0.129 | 0.878 | No |
| 4 | 1 | 1 | $Y \to X$ | 0.412 | 0.662 | No |
| | | | $X \to Y$ | 0.129 | 0.878 | No |
| 5 | 1 | 1 | $Y \to X$ | 0.513 | 0.599 | No |
| | | | $X \to Y$ | 0.164 | 0.848 | No |
| 6 | 4 | 1 | $Y \to X$ | 6.463 | 0.002 | Yes |
| | | | $X \to Y$ | 1.799 | 0.168 | No |
| 7 | 4 | 1 | $Y \to X$ | 4.035 | 0.019 | Yes |
| | | | $X \to Y$ | 16.653 | 0.000 | Yes |

*3.1.2. Correlation decision of time warping series.* Select the Sinc function with volatility ($\mathrm{Sinc}(x) = sin\pi x/\pi x,\ x \in [-6, 6]$) as the correlation research object of the simulated data, and do the following two functions: $d_1(t) = 0.01t^2 - 0.36$ and $d_2(t) = 0.005t\,(t-6)\,(t+6)$. In the two cases, the correlation coefficient varing trends with the standard Sinc function and time-lag series are shown in Fig. 3.

It can be seen from Fig. 3, right part: two time-lag series correlation coefficient curves have obvious convex phenomenon, and bounds of two correlation coefficients are [0.991,0.996] and [0.914,0.962]. Then it can be decided that the two groups of

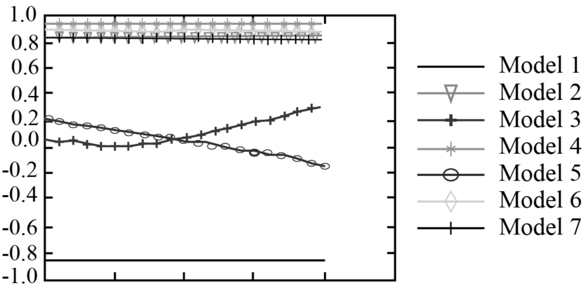


Fig. 2. Diagram of time-lag series correlation coefficient changing

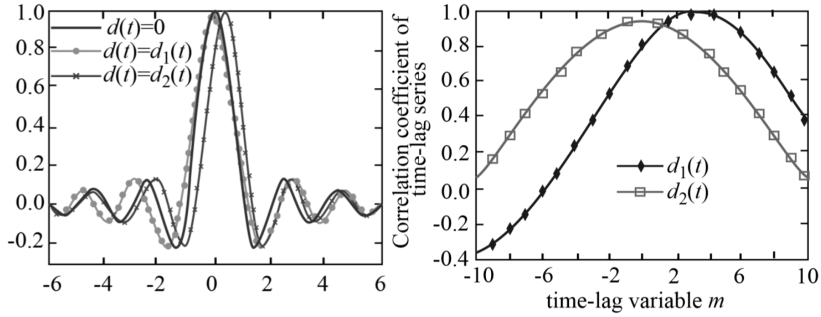


Fig. 3. Sinc function and the correlation coefficient variation of time-lag series: left–Sinc function and the function in two time differences, right–correlation coefficient variation of time-lag series under two kinds of time differences

series have correlation. And the average hysteresis amount with sequence of time-lag function of $d_1(t)$ and $d_2(t)$ and the standard Sinc sequence are 0 and 3, respectively.

### 3.2. Curve registration

This section is mainly testing the performance of the MGMF algorithm and compare with the classical CMRM algorithm [8], maximum likelihood registration (MLR) and self-modeling registration (SMR). For the sake of fairness, CMRM algorithm and MLR results are the average results of 5 experiments run. Machine configuration is: Intel Quad CPU (2.83 GHz dominant frequency), 3G memory.

The time difference functions are $d_1(t) = 0.01t^2 - 0.36$ and $d_2(t) = 0.005t \cdot (t-6)(t+6)$, respectively. Put them with the noise-containing Sinc function and the standard Sinc function to do curve registration. In the two kinds of time difference functions, the align effects of 4 kinds of alignment methods after parameter adjustment are shown in Fig. 4.

Figure 4 shows that: when the time difference function is $d_1(t)$, the MLR alignment effect is poor, and the 3 others are closer to the actual value; when the time difference function is $d_2(t)$, the MLR and CMRM alignment effect is poor, and the SMR and MGMF are very close to the time difference function.

The above experiments show that for the simple time difference function, the

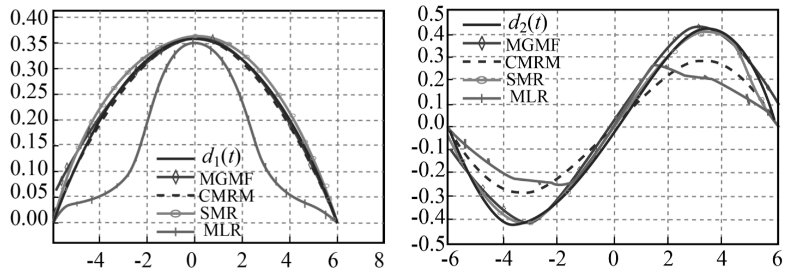


Fig. 4. Align effects of 4 kinds of methods: left–time difference function $d_1(t)$, right–time difference function $d_2(t)$

precisions of CMRM, SMR and MGMF are high, but the CMRM efficiency is poor. When the time difference function is complicated, the SMR and MGMF results are in better alignment, but the efficiency and stability of SMR is better than MGMF.

## 4. Conclusion

In this paper, the upper and lower bounds of the population correlation coefficient are given at a certain significance level and used to discriminate the correlation. Spurious regression problems arise for more reasons, and there is not yet find a rigorous and accurate identification method. In this paper, starting from the main reasons of the spurious regression, we get the characteristics of the correlation coefficient of the time shift sequence, which can eliminate most of the common spurious regression phenomenon; For another kind of correlation error, the correlation coefficient can be determined from the characteristics of the time series correlation coefficient. Based on the correlation sequence with time warping, a model based on correlation coefficient maximization and an improved MGMF algorithm are established. The applicability of the model is more extensive than the AISE criteria. The experimental results show that the correlation method in this paper is more effective than Pearson linear correlation coefficient, Spearman rank correlation coefficient. Kendall rank correlation coefficient and Granger causality test in spurious regression recognition. The proposed MGMF algorithm is significantly superior to CMRM, SMR and MLR in most cases.

**References**

[1] F. Shang, Y. Jiang, A. Xiong, W. Su, L. He: *A node localization algorithm based on multi-granularity regional division and the lagrange multiplier method in wireless sensor networks*. Sensors (Basel) *16* (2016), No. 11, e1934.

[2] J. Dai, X. Liu, F. Hu: *Research and application for grey relational analysis in multi-granularity based on normality grey number*. Scientific World Journal (2014), No. 2, ID 312645.

[3] A. Arribas-Gil, H. G. Müller: *Pairwise dynamic time warping for event data*. Computational Statistics & Data Analysis *69* (2014), 255–268.

[4] P. M. Thompson, J. Moussai, S. Zohoori, A. Goldkorn, A. A. Khan, M. S. Mega, G. W. Small, J. L. Cummings, A. W. Toga: *Cortical variability and asymmetry in normal aging and Alzheimer's disease.* Cerebral Cortex *8* (1998), No. 6, 492–509.

[5] X. Huang, N. Paragios, D. N. Metaxas: *Shape registration in implicit spaces using information theory and free form deformations.* IEEE Transactions on Pattern Analysis & Machine Intelligence *28* (2006), No. 8, 1303–1318.

[6] D. Telesca: *Bayesian analysis of curves shape variation through registration and regression.* Book: Frontiers in Probability and the Statistical Sciences, Nonparametric Bayesian Inference in Biostatistics, Springer (2015), 287–310.

[7] A. Kneip, J. O. Ramsay: *Combining registration and fitting for functional models.* Journal of the American Statistical Association *103* (2008), No. 483, 1155–1165.

[8] T. Chau, S. Young, S. Redekop: *Managing variability in the summary and comparison of gait data.* Journal of Neuroengineering & Rehabilitation *2* (2005), No. 1, paper 22.

[9] A. Kneip, X. Li, K. B. MacGibbon J. O. Ramsay: *Curve registration by local regression.* Canadian Journal of Statistics *28* (2000), No. 1, 19–29.

[10] S. A. Ahmadi, T. Sielhorst, R. Stauder, M. Horn, H. Feussner, N. Navab: *Recovery of surgical workflow without explicit models.* Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention, 1–6 October 2006, Copenhagen, Denmark, Springer Nature, LNCS *4190* (2006), Part No. 1, 420–428.

[11] K. Wang, T. Gasser: *Asymptotic and bootstrap confidence bounds for the structural average of curves.* Annals of Statistics *26*, (1998), No. 3, 972–991.

Received June 29, 2017