CHIN. PHYS. LETT. Vol. 34, No. 12 (2017) 120501

Generalized Multivariate Singular Spectrum Analysis for Nonlinear Time Series De-Noising and Prediction *

Yi Ji(吉奕)[1], Hong-Bo Xie(谢洪波)[2]**

[1]School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013
[2]ARC Centre of Excellence for Mathematical and Statistical Frontiers, Queensland University of Technology, Brisbane, QLD 4000, Australia

(Received 1 August 2017)

Singular spectrum analysis and its multivariate or multichannel singular spectrum analysis (MSSA) variant are effective methods for time series representation, denoising and prediction, with broad application in many fields. However, a key element in MSSA is singular value decomposition of a high-dimensional matrix stack of component matrices, where the spatial (structural) information among multivariate time series is lost or distorted. This vector-space model also leads to difficulties including high dimensionality, small sample size, and numerical instability when applied to multi-dimensional time series. We present a generalized multivariate singular spectrum analysis (GMSSA) method to simultaneously decompose multivariate time series into constituent components, which can overcome the limitations of conventional multivariate singular spectrum analysis. In addition, we propose a SampEn-based method to determine the dominant components in GMSSA. We demonstrate the effectiveness and efficiency of GMSSA to simultaneously de-noise multivariate time series for attractor reconstruction, and to predict both simulated and real-world multivariate noisy time series.

PACS: 05.45.Tp, 05.45.Ac, 95.75.Wx DOI: 10.1088/0256-307X/34/12/120501

Singular spectrum analysis (SSA) aims to decompose a time series into the sum of a small number of interpretable components such as a slowly varying trend, oscillatory components and 'structureless' noise.[1] Based on these components, different algorithms for de-noising, change-point detection, missing data-point interpolation, synchronization detection, feature extraction, and prediction have been developed for a range of broad fields.[2,3] Recently, SSA has been extended to multivariate or multichannel singular spectrum analysis (MSSA), to accommodate multivariate time series.[3] MSSA has achieved certain success in the analysis of several real-world multivariate time series. However, when the trajectory matrices of all variables are concatenated into one matrix, the dimension of the latter is very large. It is therefore difficult to evaluate the covariance matrix accurately due to its large size as well as the relatively small number of training samples. In addition to this 'curse of dimensionality' dilemma and 'small sample size' problem, computing the singular vectors of a large matrix is computationally expensive. Furthermore, the spatial (structural) information among multivariate time series is lost or distorted in the stacked single matrix. There are also other issues such as numerical instability in subsequent prediction or pattern classifications as well as decreasing the computational complexity and storage requirements.

In this study, we present a generalized multivariate singular spectrum analysis (GMSSA) method to simultaneously decompose multivariate time series into their constituent components, where the trajectory matrices of each time series are not required to be re-arranged into one matrix. In addition, the estimation of dominant components/eigenvalues in SSA as well as in principal component analysis (PCA) is still a known challenge, where finding the maximal gap between eigenvalues or flat noise floor represents the major existing method.[4,5] In practice, such simple criteria can either lead to information loss or overfit the underlying dynamics, especially when dealing with time series stemming from nonlinear processes.[4,5] In this study, an adaptive dominant component estimation method based on sample entropy (SampEn) is proposed and embedded in our GMSSA approach.

The generalized multivariate singular spectrum analysis approach consists of four steps: (1) embedding, (2) rank factorisation, (3) grouping, (4) signal recovery. Given a multivariate time series data set $\{T_i\}_{i=1}^{c}$, where each time series sample $T_i = \{x_i^1, x_i^2, \ldots, x_i^n\}$, each time series T_i is transformed into a trajectory matrix

$$\boldsymbol{A}_i = \begin{bmatrix} X_{i,1}^T \\ X_{i,2}^T \\ \vdots \\ X_{i,m}^T \end{bmatrix} = \begin{bmatrix} x_i^1 & x_i^2 & \cdots & x_i^d \\ x_i^2 & x_i^3 & \cdots & x_i^{d+1} \\ \vdots & \vdots & \cdots & \vdots \\ x_i^m & x_i^{m+1} & \cdots & x_i^n \end{bmatrix}, \quad (1)$$

where d is the window length, and $m = n-d+1$ represents the number of embedded delay-shift row vectors.

For the trajectory matrix set $\{\boldsymbol{A}_i\}_{i=1}^{c}$, we define the averaged horizontal and vertical covariance matri-

*Supported by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).
**Corresponding author. Email: hongbo.xie@qut.edu.au
© 2017 Chinese Physical Society and IOP Publishing Ltd

ces

$$\boldsymbol{F} = \sum_{i=1}^{c} (\boldsymbol{A}_i - \bar{\boldsymbol{A}})(\boldsymbol{A}_i - \bar{\boldsymbol{A}})^T, \tag{2}$$

$$\boldsymbol{G} = \sum_{i=1}^{c} (\boldsymbol{A}_i - \bar{\boldsymbol{A}})^T(\boldsymbol{A}_i - \bar{\boldsymbol{A}}), \tag{3}$$

where $\bar{\boldsymbol{A}} = \sum_i \boldsymbol{A}_i/c$. Let $\boldsymbol{U}$ and $\boldsymbol{V}$ be the eigenvector matrices of $\boldsymbol{F}$ and $\boldsymbol{G}$, such that

$$\boldsymbol{F} = \sum_{l=1}^{m} \lambda_l u_l u_l^T, \boldsymbol{U} \equiv (\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_m), \tag{4}$$

$$\boldsymbol{G} = \sum_{r=1}^{d} \zeta_r v_r v_r^T, \boldsymbol{V} \equiv (\boldsymbol{v}_1, \mathbf{v}_2, \ldots, \boldsymbol{v}_d). \tag{5}$$

We can then define

$$\begin{aligned} \boldsymbol{M}_{i,j} =& \boldsymbol{U}^T \boldsymbol{A}_i \boldsymbol{V}(:,j) = \boldsymbol{U}^T \boldsymbol{A}_i \boldsymbol{v}_j, \\ & (i = 1, \ldots, c; j = 1, \ldots, d), \end{aligned} \tag{6}$$

$$\begin{aligned} \tilde{\boldsymbol{A}}_{i,j} =& \boldsymbol{U}\boldsymbol{M}_{i,j}\boldsymbol{V}^T(:,j) = \boldsymbol{U}\boldsymbol{M}_{i,j}\boldsymbol{v}_j^T, \\ & (i = 1, \ldots, c; j = 1, \ldots, d). \end{aligned} \tag{7}$$

With such a transformation, each matrix $\boldsymbol{A}_i (i = 1, \ldots, c)$ is deconstructed into the sum of d rank-one matrices $\tilde{\boldsymbol{A}}_{i,j} (j = 1, \ldots, d)$.

The grouping step splits the above elementary rank-one matrices $\tilde{\boldsymbol{A}}_{i,j}$ into several groups, summing the matrices within each group. Let $L = \{l_1, l_2, \ldots, l_p\}$ be a group of indices $l_1, l_2, \ldots, l_p$. Then matrix $\tilde{\boldsymbol{A}}_{i,L}$ corresponding to the group L is defined as $\tilde{\boldsymbol{A}}_{i,L} = \tilde{\boldsymbol{A}}_{i,l_1} + \tilde{\boldsymbol{A}}_{i,l_2} + \ldots + \tilde{\boldsymbol{A}}_{i,l_p}$. The split of the set of indices $\{1, 2, \ldots, d\}$ into disjoint subsets $L_1, L_2, \ldots, L_s$ corresponds to the representation

$$\tilde{\boldsymbol{A}}_i = \tilde{\boldsymbol{A}}_{i,L_1} + \tilde{\boldsymbol{A}}_{i,L_2} + \ldots + \tilde{\boldsymbol{A}}_{i,L_s}, \ (i = 1, \ldots, c). \tag{8}$$

The procedure of choosing the sets $L_1, L_2, \ldots, L_s$ to group matrices $\tilde{\boldsymbol{A}}_{i,j}$ is termed the grouping. If the original series contains signal and noise, we then consider two groups of indices $L_1 = \{1, 2, \ldots, k\}$ and $L_2 = \{k+1, \ldots, d\}$, and associate the group L_1 with the signal component, and the group L_2 with the noise. In conventional SSA/MSSA/PCA, the ranked eigenvalues are employed to group the components based on their values, which does not consider the individual time series characteristics. If the selected number of eigenvalues is large, the system dynamics is overfitted with noise. On the other hand, if the number of the eigenvalues is low, the system dynamics will not be well modeled enough, and there will be information loss. SampEn is a nonlinear measure to quantify the degree of regularity or complexity in time series.[6–8] It can be roughly defined as the negative average natural logarithm of the conditional probability that two sequences are similar for d points within the tolerance remain similar at the next point. In this study, we propose to split the set of indices $\{1, 2, \ldots, d\}$ for each $\tilde{\boldsymbol{A}}_i$ based on the SampEn of $\tilde{\boldsymbol{A}}_{i,j}$. A modified version of SampEn, showing better relative consistency and more robustness to noise,[8] is employed in this study. The aim of SSA is to decompose the original series into the sum of a small number of independent and interpretable components such as a slowly varying trend, oscillatory components and structureless noise. It is well known that larger SampEn values correspond to broader banded spectra, whilst lower SampEn values correspond to more peaked spectra.[9] Therefore, it is natural to group $\tilde{\boldsymbol{A}}_{i,j}$ with large SampEn into noise, whilst low SampEn values correspond to trend and medium values correspond to the oscillatory components. For real-world multivariate time series, the group number of each original time series depends on the physical characteristics of the sub-system which generates the individual time series. Using the SampEn to discriminate the embedding matrix complexity, the group number may be varied for each $\tilde{\boldsymbol{A}}_i$, i.e., $L_1, L_2, \ldots, L_s$, which will differ for the individual time series. In contrast to the total variance or maximal eigenvalue gap criteria in existing techniques, multivariate time series utilize the same cut-off eigenvalue.

The fourth and last step of GMSSA converts each regrouped matrix in Eq. (8) back to the initial time series of length n. Diagonal averaging is employed to generate reconstructed elements of each resultant matrix by averaging over positive slope skew diagonal groupings. The new element has the same position as the corresponding elements in the original series. Consider matrix $\tilde{\boldsymbol{A}}_i$ with elements $y_{\mathrm{gh}}, 1 \leqslant g \leqslant m, 1 \leqslant h \leqslant d$. Let $d^* = \min(m, d)$, $m^* = \max(m, d)$ and $n = m + d - 1$. Let $y_{ij}^* = y_{ij}$ if $m < d$ and $y_{ij}^* = y_{ji}$ otherwise. Diagonal averaging transfers matrix $\tilde{\boldsymbol{A}}_i$ to a series $\tilde{T}_i = \{\tilde{x}_i^1, \tilde{x}_i^2, \ldots, \tilde{x}_i^n\}$ according to

$$\tilde{x}_i^t = \begin{cases} \frac{1}{t} \sum\limits_{q=1}^{t} y_{q,t-q+1}^*, & 1 \leqslant t < d^*, \\ \frac{1}{d^*} \sum\limits_{q=1}^{d^*} y_{q,t-q+1}^*, & d^* \leqslant t \leqslant m^*, \\ \frac{1}{n-t+1} \sum\limits_{q=t-m^*+1}^{n-m^*+1} y_{q,t-q+1}^*, & m^* < t \leqslant n, \end{cases} \tag{9}$$

which corresponds to the averaging of the matrix elements over the 'diagonals' $i + j = t + 1$. Such diagonal averaging, applied to any multivariate submatrix $\tilde{\boldsymbol{A}}_{i,j}$ in Eq. (8), generates an n length time series, and thus the original series T_i is decomposed into the sum of s series

$$T_i = \tilde{T}_{i,1} + \tilde{T}_{i,2} + \ldots + \tilde{T}_{i,s}, \quad i = 1, 2, \ldots, c. \tag{10}$$

To demonstrate the usefulness of the proposed MGSSA method, we first apply it to reconstruct a benchmark dynamic system embedded in noise: the

chaotic Lorenz system. The chaotic Lorenz system with three variables x, y, z and its parameters are defined in Ref. [1]. The integration step is set to 0.015, and the first 5000 points are discarded to eliminate initial transients.

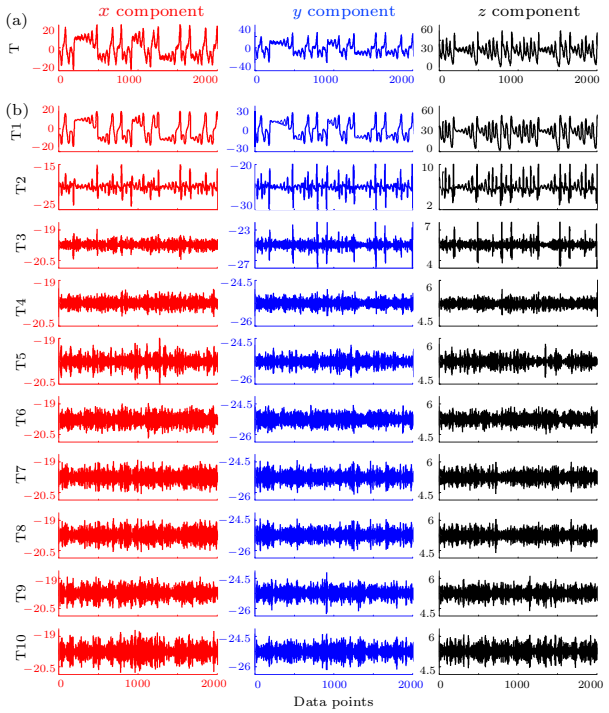

Fig. 1. (a) Three state variables (x, y and z) of noisy Lorenz series with 10% NL. (b) Ten reconstructed components (T1–T10) of each variable using GMSSA.

Noise is then superimposed onto the time series through the addition of independent and identically distributed (i.i.d.) Gaussian white noise at various noise levels (NL). The embedding dimension is set to ten, which is the maximal dimension of individual variables x, y and z obtained using false nearest neighbors.[10] Figure 1 illustrates the GMSSA decomposition for the noisy Lorenz at 10% NL. The top panel of the figure shows the original noisy data, while the lower panels illustrate the ten reconstructed components of the state variables. The first component, showing the main trend, is similar to the original series.

Figures 2 illustrates the SampEn of the same Lorenz series, but with various noise levels ranging from 2% to 10% in steps of 2%. The tolerance parameter r is set to 0.2 for estimation of the SampEn. SampEns on all panels from (a) to (c) correspond in the order from the leading to subordinate eigenvalues. For each sub-panel, the SampEn initially increases, corresponding to the dominant components from the low to high frequency band. SampEns then fluctuate severely for various noise levels, indicating the noise components in each variable. The SampEn for both Gaussian and uniform distribution noise is known to be over 1.25 when tolerance r is 0.2.[11] Therefore, we select those components with SampEn less than 1.25 in the ascending phase of each curve as the dominant components. In conventional SSA and MSSA analysis, identification of the dominant components containing the most information on the system is often based on visual inspection of the eigenvalue plot without any quantitive criteria. In this study, based on the SampEn criterion, we can accurately deduce that the first three components contain the most information in noisy Lorenz system.

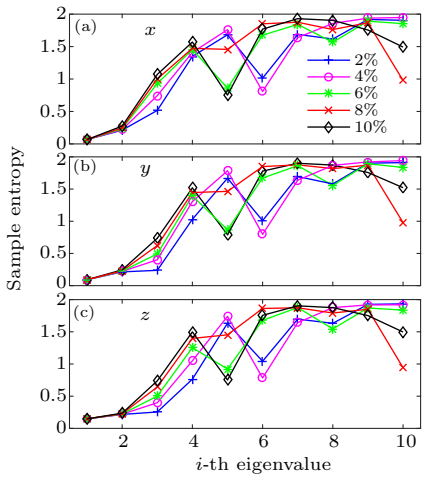

Fig. 2. Sample entropy distribution of noisy Lorenz variables x (a), y (b) and z (c) with noise levels varying from 2% to 10% in steps of 2%.

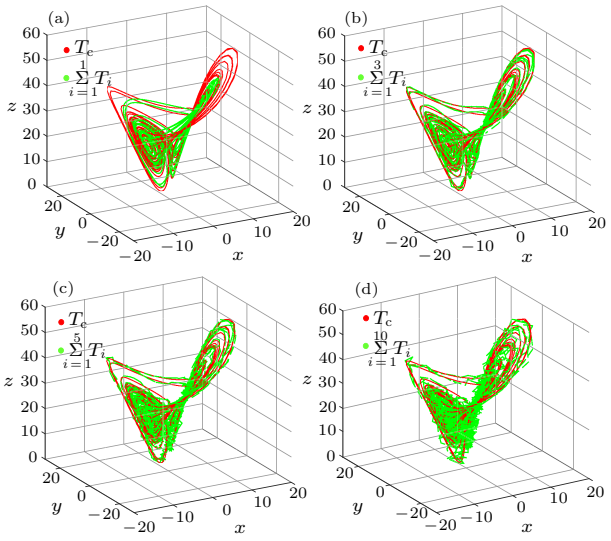

Fig. 3. The 3D reconstructed attractor for the clean Lorenz series T_c (red) and the sum of first one (a), first three (b), first five (c), and all ten (d) $\tilde{T}_{i,s}$ components obtained from GMSSA (green) with 10% Gaussian noise.

Figure 3 shows the reconstructed attractor of the clean Lorenz (red), along with the reconstructed attractors from the sum of various components with added 10% noise level (green), in which the first sub-panel shows the attractor of the first component. From visual inspection, the clean attractor's envelope (red) exceeds that of the first component (green) in

sub-panel A, indicating an information loss when using only the first component to reconstruct the noisy attractor. Sub-panels C and D of Fig. 3 show the reconstructed attractors formed from the sum of the first five and ten $\tilde{T}_{i,s}$ components, respectively. We observe that the green envelopes of these two sub-panels exceed that of the clean data (red). This implies that redundant components do not contribute to the accuracy of reconstructed attractors due to the noise component they represent. However, the clean data attractor is best matched to the sum of the first three components shown in sub-panel B. This is consistent with the fact that the first three components contain the most essential information of Lorenz system using the SampEn criterion indicated in Fig. 2. It can therefore be concluded that the dynamic trajectory of the noisy Lorenz dynamic system can be recovered by selecting an appropriate number of components obtained from GMSSA using the SampEn criterion.

We also consider the issue of predicting noisy multivariate time series using GMSSA. Currently, the existing SSA and MSSA approaches can only be applied to predict those time series that approximately satisfy the linear recurrent formulae,[2] which are violated by most nonlinear time series. On the other hand, SSA has been applied as an efficient pre-processing algorithm to eliminate time series noise, which results in modified representation of input vectors for prediction. Using such SSA pre-processed data as the input, there is considerable improvement in prediction accuracy using support vector machine (SVM) or artificial neural network (ANN) approaches.[12] However, these SSA-SVM or SSA-ANN prediction models are limited to univariate time series. In fact, the time series observed from complex systems generally consists of multiple variables, with complementary information of the underlying dynamic system contained in multivariate time series.[13] It has been proved that a prediction model with multivariate time series can achieve higher accuracy than those with univariate time series.[14] However, the hybrid model to predict the multivariate time series is still lacking, i.e., using pre-processed multiple time series as the input to predict a univariate output. In this study, we present the hybrid GMSSA-SVM approach for multivariate time series prediction, in which GMSSA smooths the noisy data and SVM fulfills the robust regression analysis. To construct the prediction model input–output pairs, the input set consists of concatenating the vectors $[x_i^e, x_i^{e+1}, \ldots x_i^{e+d-1}]_{i=1}^c$ in the same row for $\{\tilde{\boldsymbol{A}}_i\}_{i=1}^c$ (see Eq. (8)). The corresponding output consists of x_i^{e+h} for time series T_i, where h is the desired prediction horizon. The prediction is therefore based on a direct prediction model rather than the iterative scheme. For the Lorenz series, only the row vectors obtained from the sum of the first three elementary matrices, i.e., $\tilde{\boldsymbol{A}}_{i,j}(i = 1, 2, \ldots, c; j = 1, 2, 3)$ are used as the input to construct the prediction model. The input vector dimensions are therefore 30. The output of the model is the x-variable of the Lorenz series. The training set consists of 1500 points, while the test set consists of 500 points for evaluation of the prediction performance. Prediction quality was evaluated in terms of the mean squared error (MSE) and the correlation coefficient (CC). Figure 4 shows the MSEs and CCs from one to twenty-step predictions for the Lorenz series with 10% NLs. The corresponding MSEs of SVM models without GMSSA preprocessing are also evaluated for comparison. With increasing the prediction lead time, there is a clear trend of increased MSE and decreased CC for both methods. However, the MSE for GMSSA-SVM is significantly lower than SVM for all the prediction time leads, whilst the CC for GMSSA-SVM is persistently larger than SVM. The results clearly demonstrate that GMSSA-SVM provides an efficient approach to improve multivariate time series prediction accuracy.

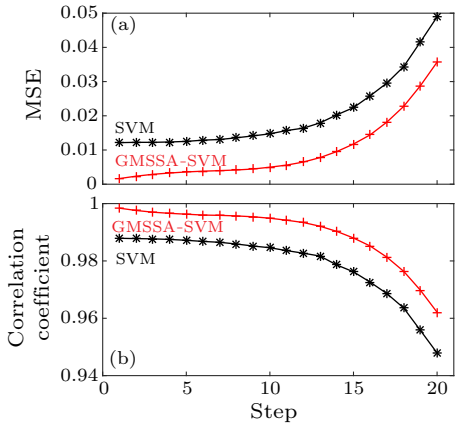

Fig. 4. The mean squared error and correlation coefficient of multi-step prediction for the Lorenz series with 10% NL.

Embedding dimension is an important parameter in SSA approaches. To expand the system dynamics as much as possible and to indicate the variation of SampEn between components, we set the dimension as the maximal dimension of individual variables. We have compared the performance of GMSSA with dimension varying from six to ten and there is no significant difference found.

To test the applicability of GMSSA to real life data, we apply GMSSA to a financial data set, namely, the Shanghai stock exchange (SSE) composite index. We combine closing price, opening price, high price, and low price time series into one multivariate time series to predict the evolution of the closing price. The data length is 2000 daily points starting from 26/09/2005. The first 1500 points are used for training and the remaining for testing prediction. The SampEn distribution of four series indicates that the number of dominant components is 2. Table 1 summarizes the MSE and CC for the prediction time from one to five days, which further indicates the superiority of the hybrid GMSSA method for real life data prediction.

Table 1. MSE and CC for prediction of the Shanghai stock exchange composite index.

	Step	1	2	3	4	5
MSE	GMSSA-SVM	5.102×10^{-4}	6.375×10^{-4}	9.213×10^{-4}	1.412×10^{-3}	2.061×10^{-3}
	SVM	3.140×10^{-4}	1.061×10^{-3}	1.748×10^{-3}	2.398×10^{-3}	3.154×10^{-3}
CC	GMSSA-SVM	0.972	0.964	0.948	0.922	0.889
	SVM	0.983	0.941	0.904	0.870	0.832

We have presented the generalized multivariate singular spectrum analysis framework which can simultaneously decompose multiple time series into their respective constituent components, including the slowly varying trend, oscillatory components and the unstructured noise. The curse of dimensionality dilemma diminishes in GMSSA, since GMSSA operates on the trajectory matrix of each time series rather than a high-dimensional matrix stack of all trajectory matrices. In addition, since the input feature vectors analyzed are actually the row vectors of the training samples, the feature set is significantly enlarged. Therefore, the small sample size problem is not an issue in GMSSA. In terms of these two advantages, it is evident that the covariance matrix in GMSSA can be estimated more robustly and accurately than MSSA. Moreover, the two-dimensional spatial information is well preserved using the original matrix. Determining the optimal number of components is a fundamental problem in component learning or modeling. In the present study, we have presented a SampEn criterion to estimate the dominant components, which can be employed in any kind of SSA, multivariate SSA, or other component analysis techniques. Here we have used GMSSA to reconstruct the attractors of noisy Lorenz dynamic systems. We have also used the dominant components of GMSSA to predict both simulated and real life noisy time series. In this work, only white noise is considered. However, the proposed algorithm can be modified with extra whitening pre-processing and de-whitening pro-processing to remove colored noise in the input multivariate time series.[15]

References

[1] Xie H B, Guo T, Sivakumar B, Liew A W C and Dokos S 2014 *Proc. R. Soc.* A **470** 20140409
[2] Hassani H, Heravi S and Zhigljavsky A 2013 *J. Forecasting* **32** 395
[3] Hossein H 2013 *Int. J. Energy Stat.* **1** 55
[4] Song J J, Ren Y and Yan F L 2009 *Comput. Biol. Chem.* **33** 408
[5] Pilgram B and Schappacher W 1998 *Int. J. Bifurcation Chaos Appl. Sci. Eng.* **8** 571
[6] Xie H B, Guo J Y and Zheng Y P 2010 *Phys. Lett.* A **374** 3926
[7] Xie H B, He W X and Liu H 2008 *Phys. Lett.* A **372** 7140
[8] Xie H B, Chen W T, He W X and Liu H 2011 *Appl. Soft Comput.* **11** 2871
[9] Xie H B, Guo J Y and Zheng Y P 2010 *Ann. Biomed. Eng.* **38** 1483
[10] Kennel M B, Brown R and Abarbanel H D I 1992 *Phys. Rev.* A **45** 3403
[11] Jiang J and Xie H B 2016 *Chin. Phys. Lett.* **33** 100501
[12] Wu C L and Chau K W 2013 *Eng. Appl. Artif. Intel.* **26** 997
[13] Wang X Y and Han M 2015 *Eng. Appl. Artif. Intel.* **40** 28
[14] du Preez J and Witt S F 2003 *Int. J. Forecasting* **19** 435
[15] Xie H B, Sivakumar B, Boontra T W and Mengersen K 2017 *IEEE Trans. Fuzzy Syst.* (in press)