Improved time series prediction with a new method for selection of model parameters

2006 J. Phys. A: Math. Gen. 39 L483

(http://iopscience.iop.org/0305-4470/39/30/L01)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **39** (2006) L483–L491 doi:10.1088/0305-4470/39/30/L01

LETTER TO THE EDITOR

Improved time series prediction with a new method for selection of model parameters

A M Jade[1], **V K Jayaraman and B D Kulkarni**

Chemical Engineering and Process Development Division, National Chemical Laboratory, Pune 411008, India

E-mail: bd.kulkarni@ncl.res.in

Received 24 February 2006, in final form 21 June 2006
Published 12 July 2006
Online at stacks.iop.org/JPhysA/39/L483

Abstract

A new method for model selection in prediction of time series is proposed. Apart from the conventional criterion of minimizing RMS error, the method also minimizes the error on the distribution of singularities, evaluated through the local Hölder estimates and its probability density spectrum. Predictions of two simulated and one real time series have been done using kernel principal component regression (KPCR) and model parameters of KPCR have been selected employing the proposed as well as the conventional method. Results obtained demonstrate that the proposed method takes into account the sharp changes in a time series and improves the generalization capability of the KPCR model for better prediction of the unseen test data.

PACS numbers: 05.45.Tp, 05.45.Df, 07.05.Mh

1. Introduction

Several methods have been proposed in the literature for prediction of time series data [1–15]. The most familiar approaches include the linear methods such as ARX, ARMA, etc and the nonlinear methods such as algorithms based on artificial neural networks [6–11]. Recently kernel-based machine learning tools such as support vector regression (SVR) and kernel principal component regression (KPCR) have become very popular because of their state-of-the-art performance [12–15]. All these methods split the data into three disjoint sets, namely, training, validation and test sets. Subsequently, the model parameters in the algorithm are optimized by minimizing the root mean square error of the predicted validation set. Finally, the performance is gauged by the test error. For time series data possessing sharp changes, the selection of model parameters based only on the criterion of RMS error may produce higher

[1] Present address: Analytics Centre of Excellence, Persistent Systems Pvt. Ltd, Pingala & Aryabhat, Erandwana, Pune 411004, India.

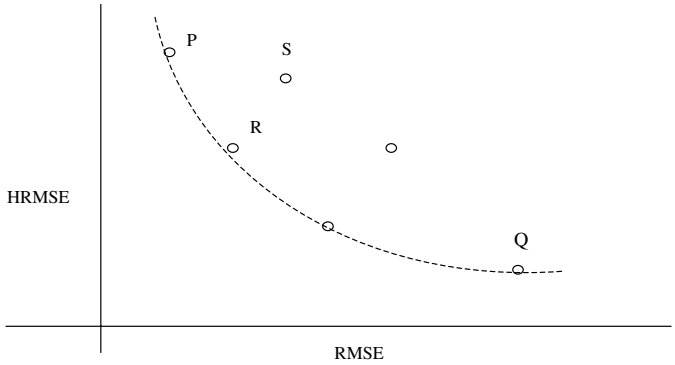

Figure 1. Concept of Pareto-optimal solutions in multiobjective optimization.

generalization errors. There is therefore a need for a robust measure, which will take account of these sharp changes or singularities occurring in a time series. A methodology that picks up the local scaling behaviour of the time series would be able to readily reveal such singularities.

In this letter we present the wavelet transform modulus maxima (WTMM)-based method for characterizing and quantifying the singularities in a chaotic time series [16]. The method provides the density estimates of the local Hölder exponents that characterize the regular/irregular local behaviour of time series. Higher the value of the local Hölder exponent, the more regular is the local behaviour of time series and vice versa. The density estimates of the local Hölder exponents represent the most informative features regarding the singularities in the time series. Thus an estimate of the error in the density spectrum of the predicted validation set, which we will henceforth denote as HRSME, can be very useful for tuning the model parameters. For certain time series data it may be possible that both RMSE and HRMSE information would be useful for obtaining optimal performance. In this work, errors in the density estimates (HRMSE) (along with regular RMSE) of the validation set have been employed as an additional criterion for the selection of optimal model parameters. Thus the problem of model selection is formulated in terms of a multiobjective optimization i.e. the selection of model parameters has been done by minimizing both criteria, namely, RMS errors based on the original time series as well as on the error in the singularity distribution. In this problem, we have to find the decision vector (parameters of model used for time series prediction), which will minimize RMSE as well as HRMSE. Multiobjective optimization, however, gives rise to a set of optimal solutions, instead of a single one [17]. These optimal solutions are called as Pareto-optimal solutions. This concept can be illustrated employing figure 1, which depicts a plot of RMSE versus HRMSE. The point 'P' represents the solution with minimum RMSE but has higher HRMSE and point 'Q' represents least HRMSE but high RMSE. Since both objectives are equally important one cannot say that solution 'P' is better than 'Q' or vice versa. All such solutions (marked by the dashed line) are Pareto-optimal solutions. Also in the figure, there are few points (e.g. 'S') which are not members of the Pareto set. It can be seen that solution 'R' in the decision space has lower RMS and HRMS errors and hence is better than solution 'S' considering both the objectives. Thus solutions like 'S' are known as dominated or inferior solutions and solutions like 'R' belonging to the Pareto-optimal set are often known as non-dominated solutions. Also it is clear that no solution in the Pareto-optimal set is better or worse than the other considering both the objectives (RMSE and HRMSE). In the present study we have used this concept of non-dominated Pareto-optimal solutions for finding the optimal parameters of the KPCR model to improve the generalization

capability of the model. Kernel PCA, a nonlinear version of PCA, has recently been extensively used because of its computational simplicity and nonlinear feature extraction and denoising capabilities [14, 15, 18]. We have chosen KPCR because of its excellent performance on time series prediction problems [14, 15]. Moreover, only two parameters are needed to be tuned for model selection of KPCR. The efficacy of the proposed method has been tested on two simulated chaotic time series and one time series based on real observations. This letter is organized as follows: in section 2, we have described the key steps involved in the proposed algorithm. Section 3 includes the discussion on characterizations of the singularities and their analysis based on the wavelet transform modulus maxima (WTMM) method. In section 4, we have illustrated KPCR for time series prediction. Section 5 includes the case studies used for time series prediction. Section 6 comprises the results and discussions and section 7 provides salient conclusions of the work.

2. Proposed algorithm for KPCR model selection

The key steps involved in the proposed algorithm are described below:

(1) Divide the available time series data into three segments namely training, validation and test.
(2) For various model parameters
   (a) Build up the KPCR model using the training data as described in section 4.
   (b) Predict the validation time series using the model.
   (c) Estimate the RMSE and HRMSE for validation time series employing the method described in section 3.
(3) Get the Pareto-optimal solutions using RMSE and HRMSE criteria.
(4) Estimate the test error using the Pareto-optimal model parameter set.

3. Singularity analysis using WTMM

The distribution of singularities and the singularity spectrum can be obtained from the well-known wavelet transform modulus maxima (WTMM)-based multifractal formalism [19–21]. The method offers global estimates of scaling properties for the characterization of a multifractal time series. The spectrum of the singularities as described in [19–21] poses certain problems of stability when applied to observational data [22]. Recently, Struzik [16] has presented a stable method for evaluating the estimation of local singularity strengths. In his methodology for estimating the local Hölder exponents, he has modelled the singularities as if they were created through a multiplicative cascading process. The method has been described in brief here and for more details readers may refer to [16, 22–26].

It can be shown that by employing the multiplicative cascade model the estimate of local Hölder exponent, $\hat{h}(x_0, s)$ at the singularity x_0 and scale s can be evaluated as [16, 23]

$$\hat{h}(x_0, s) = \frac{\log(|T_\psi[f](x_0, s)|) - (\bar{h}\log(s) + c)}{\log(s) - \log(s_N)}, \tag{1}$$

where s_N is the maximum available scale and $T_\psi[f](x_0, s)$ is the maxima of wavelet coefficient at location x_0 and at scale s. In our work we have used Mexican hat wavelet, which is the second derivative of the Gaussian function.

The mean Hölder exponent $\bar{h}$ in equation (1) can be estimated as a linear fit of the following equation:

$$\log[M(s)] = \bar{h}\log(s) + c, \tag{2}$$

where the function $M(s)$ is obtained from the partition functions

$$M(s) = \sqrt{\frac{Z(s,2)}{Z(s,0)}}. \tag{3}$$

The partition function $Z(s,2)$ can be calculated as the sum of squares of the maxima of $|T_\psi[f](x_0,s)|$ at the scale s and $Z(s,0)$ is the number of maxima at scale s.

The local Hölder estimates (from equation (1)) and their density spectrum can be estimated for original and predicted time series. The error on the distribution of these estimates can be used as criterion for model selection of time series prediction problems using KPCR.

4. Kernel principal component regression

KPCR has been chosen as a nonlinear regression method because of its successful applications in time series prediction [14, 15]. Moreover, KPCR requires only two parameters to be tuned for its model selection. Kernel principal component analysis (kernel PCA) corresponds to linear PCA in a higher dimensional feature space, which is nonlinearly related to the input space. The input data $\mathbf{x}$ are first mapped through some appropriate nonlinear function $\Phi(\mathbf{x})$. The problem formulation is in terms of dot product of the input data in the feature space that can be substituted by a kernel function. Thus *a priori* defined kernel function is used to deal with the very high-dimensional space and the dot product of transformed input vectors can be computed in the input space itself [27]. Kernel PCA and PCR have been extensively used for the purpose of nonlinear feature extraction and denoising [14, 15, 18]. KPCR has a definite advantage in dealing with multicollinearity and noise and allows us more flexibility in retaining principal components to capture the underlying nonlinear features where data observations are more than the dimensionality [14]. The method has been described in brief here and for details readers may refer to [14, 15, 18].

Consider a set of M centred input variables (regressors), $\{\mathbf{x}_k\}_1^M \in \mathbf{R}^N$ and output (response) variables $\{\mathbf{y}_k\}_1^M \in \mathbf{R}^l$.

The kernel principal component regression model for the prediction of response variable from any input vector, $\mathbf{x}$, can be expressed as [14]

$$f(\mathbf{x}) = \sum_{k=1}^{p} w_k \sum_{i=1}^{M} \alpha_i^k K(\mathbf{x}_i, \mathbf{x}) + b, \tag{4}$$

where p is the number of principal components retained in the KPCR model ($p \leqslant M$) and $K(\mathbf{x}_i, \mathbf{x})$ can be estimated using the kernel function. The variables α_i^k are computed by diagonalization of kernel matrix of the input variables and w_k are the least-squares estimates of regression coefficients. For the centralized regression model bias, b is zero. In our study we have employed Gaussian kernel, which is defined as

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{L}\right),$$

where L is the width of the Gaussian function. The number of principal components retained, p and the width parameter, L, are the two parameters that need to be tuned in fixing the KPCR model.

5. Case studies

In our analysis we have considered the three time series, two simulated and one based on real observations.

5.1. Simulated time series

In our simulations we have chosen two important benchmarking time series examples, namely, the Lorenz system [28] and Mackey–Glass [29] governed by the following set of equations.

5.1.1. The Lorenz system.

$$\frac{dx}{dt} = -\sigma x + \sigma y, \qquad \frac{dy}{dt} = Rx - y - xz, \qquad \frac{dz}{dt} = -bz + xy,$$

where $\sigma = 10$, $R = 28$, $b = 8/3$.

The time series data have been generated by integrating sets of equations using a standard Runge–Kutta routine with a step size of 0.01. The training set consisted of 500 delay vectors, formed by using an embedding dimension of 3 and a time delay of 16 for Lorenz series. The validation and test set consisted of similarly embedded 300 and 2000 delay vectors respectively.

5.1.2. Mackey–Glass.

$$\frac{dx(t)}{dt} = \frac{ax(t-\tau)}{1+x^{10}(t-\tau)} - bx(t),$$

where $a = 0.2$, $b = 0.1$, $\tau = 17$. We predict $x(t+6)$ using the input variables $x(t)$, $x(t-6)$, $x(t-12)$ and $x(t-18)$, respectively [30]. We have used training, validation and test set of size 500 each.

5.2. Laser data

We have studied the prediction of real infrared NH_3 laser data [31], contributed by U Hübner to the Sante-Fe Institute prediction contest. This data set contains two sets, consisting of 1000 and 10 000 points (http://www-psych.stanford.edu/~andreas/Time-Series/SantaFe.html). We have used embedding dimension of 4, and chosen the delay of 2 [32]. First 1000 points of first set are used for training and the first 400 points of second set are used for validation and the following 8000 points of the second set are used for test.

6. Results and discussions

We have used KPCR for time series prediction of the case studies described in section 4. Different models can be generated using various free parameters of KPCR algorithm. For KPCR we need to tune the two parameters, namely, the number of principal components retained and width parameter in Gaussian kernel. For studying the effect of model selection criteria on the performance of time series prediction, we have divided the given time series into three parts, namely, training, validation and test. Simulations can be performed with different model parameters employing the training data, and the set of parameters that predicts the least error on validation data will be selected as the optimal set and further can be employed for the prediction of the unseen test data. Generally, the RMS error on the validation data is used as an objective for selecting optimal model parameters. In other words, the parameters that yield the least RMS error on validation data are used for predicting unseen time series. In this study, we have proposed one additional criterion for the selection of optimal set of parameters based on the distribution of the local Hölder estimates. The method of evaluating the proposed criterion is as follows: (i) for each set of model parameters, estimates of the local Hölder exponents are evaluated for both original time series and the predicted time series of validation data using the algorithm described in section 3. (ii) The probability density spectra were then obtained

Table 1. Non-dominated solutions.

Number of principal component	Kernel parameter	Validation RMSE	Validation HRMSE	Test error
(a) Lorenz time series				
180	*1.0000*	*0.0414*	*0.0237*	*0.0696*
193	1.0400	0.0357	0.0277	0.0782
193	**1.0800**	**0.0342**	**0.0282**	**0.0835**
193	1.1000	0.0343	0.0279	0.0852
170	**1.3200**	**0.0439**	**0.0138**	**0.0614**
170	1.3400	0.0439	0.0192	0.0623
(b) Mackey-Glass time series				
343	**0.94**	**0.000 473**	**0.065 422**	**0.000 19**
379	0.94	0.000 517	0.039 667	0.000 185
343	0.96	0.000 474	0.062 189	0.000 188
378	0.98	0.000 532	0.039 406	0.000 187
379	1.0	0.000 535	0.038 293	0.000 181
381	1.0	0.000 538	0.032 960	0.000 182
381	**1.02**	**0.000 538**	**0.031 127**	**0.000 179**
347	*1.10*	*0.000 495*	*0.050 874*	*0.000 173*
(c) Laser time series				
330	**0.21**	**8.4848**	**0.3002**	**10.7952**
331	0.21	8.4856	0.2891	10.7954
320	0.25	8.4881	0.2523	10.3709
316	0.27	8.4912	0.1098	10.1842
317	0.27	8.4992	0.1025	10.1901
318	*0.27*	*8.4982*	*0.1030*	*10.1890*
261	0.31	8.5920	0.0997	10.0809
262	0.31	8.6037	0.0758	10.0494
262	0.33	8.5996	0.0805	9.9509
263	0.33	8.5962	0.0821	9.9542
263	0.35	8.5952	0.0840	9.8467
264	0.35	8.5951	0.0884	9.8450
281	**0.41**	**8.6717**	**0.0753**	**9.6211**

and the root mean square error between the densities of the Hölder exponents (HRMS error) is found. For the least HRMS error, the singularity distribution of predicted time series is closest to the singularity distribution of actual time series. Thus HRMS error can be used as an objective for optimizing the model parameters of time series with sharp changes. In our study, we have used both criteria of RMS and HRMS error and solved the problem of obtaining the optimal parameters as a multiobjective optimization. We have varied the two parameters of KPCR, namely, parameter (L) in Gaussian kernel and p (the number of principal components retained). Simulations were conducted with more than 10 000 sets of the model parameters. From the solutions thus obtained, we get a set of Pareto-optimal solutions as described in the first section. After getting the Pareto set, a test error for each set of parameters in this set is estimated.

Pareto sets obtained for Lorenz, Mackey–Glass and laser time series are shown in table 1. In the table, the results marked in bold are obtained by using one of the objectives (either RMSE or HRMSE). As can be seen from table 1, the model parameters ($p = 193$ and $L = 1.08$) have produced the least RMS error of 0.0342 on validation data. Using only

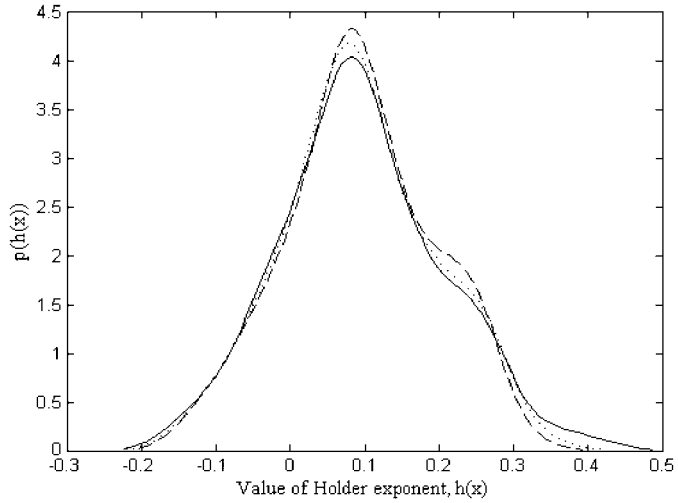

Figure 2. The distribution of the local Hölder exponents for the predicted and actual laser time series (validation set) using KPCR models. Distribution of the Hölder exponents of the actual time series is denoted by solid line and the distribution of predicted time series with a model selected by conventional criterion and singularity-based criterion alone is denoted by dashed and dotted line, respectively.

RMS error as a criterion, one may choose these parameters as optimal parameters and directly employ it for predicting test data. These parameters lead to a test error of 0.0835. For Lorenz series minimum HRMS error for validation data is obtained for the set of parameters ($p = 170$ and $L = 1.32$), which results in a test error of 0.0614. Thus use of HRMS error as a criterion has led to an improvement of 26.47% in predicting the unseen test data. A similar trend is observed in the case of laser and Mackey–Glass time series. Employing the criterion of RMS error the set of parameters ($p = 330$ and $L = 0.2100$) has been selected for laser time series which yields a test error of 10.7952, whereas optimal parameters ($p = 281$ and $L = 0.4100$) selected on the basis of new criterion result in a test error of 9.6211. Thus application of HRMS error criterion has shown an improvement of 10.88% over the conventional criterion of RMS error in predicting the test data of laser time series. Similarly an improvement of 5.78% is obtained by the application of the proposed criterion in predicting unseen Mackey–Glass time series.

The reason for the superior performance can be better understood from figure 2. We have shown in this figure, the distribution of local Hölder exponents for the predicted and actual laser time series (validation data) for KPCR models selected by conventional- and singularity-based criterion. As can be seen from the figure, the distribution of Hölder exponents for the predicted time series is closer to the distribution of the same for the actual time series in the case of the model selected by the singularity-based criterion than the one corresponding to the conventional criterion. The superior performance of the proposed method can be attributed to the fact that the model selected by minimum HRMS criterion perfectly captures local singular behaviour of the time series and thus helps in improving the generalization capability of the model.

Though employing HRMSE criterion has led to better results than the conventional criterion of RMSE in all the present case studies, there is a risk involved in choosing the model parameters based only on one of the criteria for prediction of any real time series at

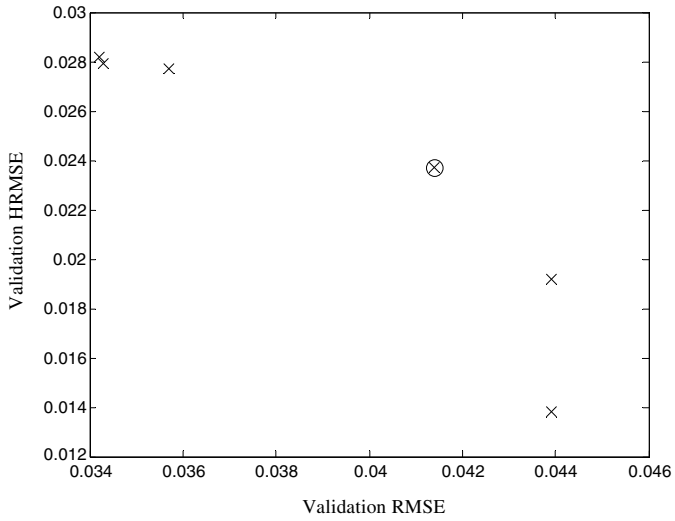

Figure 3. Non-dominated solutions for Lorenz time series (optimal point considering both objectives is marked by circle).

hand. If the range of the validation errors (both RMSE and HRMSE) for the Pareto-optimal set is narrow, then the selection of parameters using either of the criteria will not make any significant difference.

But when the range for one or both of the errors is wide, one should choose the model parameters, which are performing better in both criteria. Thus for instance for Lorenz series we will choose the parameters $p = 180$ and $L = 1$ as these parameters produce lesser RMSE as well as HRMSE. The point with the optimal parameters is marked by a circle in figure 3. By using these parameters we have obtained the test error of 0.0696 which is much lesser than the error given by the parameters using conventional criterion. In a similar way the optimal parameters performing well in both criteria are found for other case studies and are shown in italic in table 1. One can conclude from the table that the selection of parameters employing both criteria has resulted in lesser error on unseen test data than the errors produced by the parameters based on conventional criterion alone.

7. Conclusions

A new method for model selection in prediction of time series is proposed. Generally, the model parameters are selected for which the minimum RMS error is obtained on the predicted time series. This conventional criterion based only on RMS error may not take into account the sharp changes or singularities in a time series and may fail in a case where this contribution of singularities is significant. Here, we have proposed an additional criterion based on the error in the distribution of singularities in the predicted and actual time series. The distribution of singularities is evaluated through the local Hölder estimates and its probability density spectrum. Thus, the problem of model selection is solved by simultaneously minimizing both criteria, namely, RMS error based on the original time series as well as on the error in the singularity distribution. The method is tested on three time series: two simulated and one based on real observations. Predictions of these time series have been done using KPCR

and model parameters of KPCR have been selected employing the proposed as well as the conventional method. The problem now being a multiobjective optimization problem, we get a set of Pareto-optimal solutions. Results obtained demonstrate that the proposed method helps in better prediction of the unseen test data and improves the generalization capability of the KPCR model. Model selection has produced results which are better than those yielded by the conventional method in all the cases of the simulated and real time series. In general, we conclude that the new method can be very useful in the prediction of time series data having sharp singularities.

Acknowledgment

Financial assistance from the Department of Science and Technology (DST), New Delhi, India is gratefully acknowledged.

References

[1] Casdagli M 1989 *Physica* D **35** 335–56
[2] Farmer J D and Sidorowich J J 1987 *Phys. Rev. Lett.* **59** 845–8
[3] Belomestny D, Jentsch V and Schreckenberg M 2003 *J. Phys. A: Math. Gen.* **36** 11369–83
[4] Navone H D and Ceccatto H A 1995 *J. Phys. A: Math. Gen.* **28** 3381–8
[5] Ho S L and Xie M 1998 *Comput. Ind. Eng.* **35** 213–6
[6] Zhang G, Patuwo E B and Hu M Y 1998 *Int. J. Forecast.* **14** 35–62
[7] Connor J T, Martin R D and Atlas L E 1994 *IEEE Trans. Neural Netw.* **5** 240–54
[8] Rosen-Zvi M, Kanter I and Kinzel W 2003 *J. Phys. A: Math. Gen.* **36** 4543–50
[9] Elsner J B 1992 *J. Phys. A: Math. Gen.* **25** 843–50
[10] Small M and Tse C K 2002 *Phys. Rev.* E **66** 066701
[11] Freking A, Kinzel W and Kanter I 2002 *Phys. Rev.* E **65** 050903
[12] Müller K-R, Smola A J, Rätsch G, Schölkopf B, Kohlmorgen J and Vapnik V 1997 Predicting time series with support vector machines *Artificial Neural Networks—ICANN'97* ed W Gerstner, A Germond, M Hasler and J-D Nicoud (Berlin: Springer) pp 999–1004
[13] Mukherjee S, Osuna E and Girosi F 1997 *Proc. IEEE NNSP'97 (Amelia Island, FL 24–26 September 1997)*
[14] Rosipal R, Girolami M, Trejo L J and Cichocki A 2001 *Neural Comput. Appl.* **10** 231–43
[15] Jade A M, Srikanth B, Jayaraman V K, Kulkarni B D, Jog J P and Priya L 2003 *Chem. Eng. Sci.* **58** 4441–8
[16] Struzik Z R 2000 *Fractals* **8** 163–79
[17] Zitzler E, Laumanns M and Bleuler S 2004 A tutorial on evolutionary multiobjective optimization *Metaheuristics for Multiobjective Optimisation (Lecture Notes in Economics and Mathematical Systems* vol 535*)* ed X Gandibleux, M Sevaux, K Sörensen and V T'kindt (Berlin: Springer) pp 3–37
[18] Schölkopf B, Smola A J and Müller K R 1998 *Neural Comput.* **10** 1299–319
[19] Muzy J F, Bacry E and Arneodo A 1991 *Phys. Rev. Lett.* **67** 3515
[20] Muzy J F, Bacry E and Arneodo A 1993 *Phys. Rev.* E **47** 875
[21] Muzy J F, Bacry E and Arneodo A 1994 *Int. J. Bifurcation Chaos* **4** 245–302
[22] Struzik Z R 1998 CWI Report, INS-R9803
[23] Struzik Z R and Siebes A P J M 2002 *Physica* A **309** 388–402
[24] Scafetta N, Griffin L and West B J 2003 *Physica* A **328** 561–83
[25] West B J, Scafetta N, Cooke W H and Balocchi R 2004 *Ann. Biomed. Eng.* **32** 1077–87
[26] Jade A M, Jayaraman V K, Kulkarni B D, Khopkar A R, Ranade V V and Sharma A 2006 *Chem. Eng. Sci.* **61** 688–97
[27] Vapnik V 1998 *Statistical Learning Theory* (New York: Wiley)
[28] Lorenz E N 1963 *J. Atmos. Sci.* **20** 130–41
[29] Mackey M C and Glass L 1977 *Science* **197** 287–9
[30] Chen Y, Yang B and Dong J 2006 *Neurocomputing* **69** 449–65
[31] Hübner U, Abraham N B and Weiss C O 1989 *Phys. Rev.* A **40** 6354–65
[32] Bollt E M 2000 *Int. J. Bifurcation Chaos* **10** 1407–22