PHYSICAL REVIEW E **74**, 041114 (2006)

Testing for correlation structures in short-term variabilities with long-term trends of multivariate time series

Tomomichi Nakamura,[1,*] Yoshito Hirata,[2,†] and Michael Small[1,‡]
[1]*Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong*
[2]*Institute of Industrial Science, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan*
(Received 15 June 2006; published 17 October 2006)

We describe a method for identifying correlation structures in irregular fluctuations (short-term variabilities) of multivariate time series, even if they exhibit long-term trends. This method is based on the previously proposed small shuffle surrogate method. The null hypothesis addressed by this method is that there is no short-term correlation structure among data or that the irregular fluctuations are independent. The method is demonstrated for numerical data generated by known systems and applied to several experimental time series.

DOI: 10.1103/PhysRevE.74.041114 PACS number(s): 02.50.−r, 05.10.−a, 05.45.Tp

I. INTRODUCTION

In the real world systems are not always isolated from their surroundings and not always unique. Hence, we often want to know whether there is some kind of relation among them or to find similar phenomena elsewhere. This question is an old one and very important. To investigate this, time series are often the only clue. Time series usually show irregular fluctuations, and data are often modulated by long-term trends [see Figs. 1(a) and 1(b)]. When signals are similar it is not at all unusual to expect that some sort of similar activity may occur in the systems or there may be correlation structures between them [1]. On the other hand, there are cases where time series are not similar. In this case, we may have the impression that these systems are independent or have no correlation structure. However, these systems may be interconnected or interrelated in some way or another to varying degrees. There is a historical precedent for this. For example, there is a Japanese proverb: bucket makers become profitable when winds blow; yet there seems to be no relation between bucket makers and winds. This proverb means that even if some events seem to be of no relation at first glance, these systems are interconnected or interrelated [2].

A simple approach to investigate whether there is some kind of relation between two signals is to estimate the correlation coefficient. This statistic is effective to investigate similarities of long-term trends between two signals. However, it is not effective to investigate the short-term variabilities. Although it is important to know the global relation (long-term dynamics), it is also important to know the local relation (short-term dynamics). One of the useful statistics to investigate whether there are similarities in short-term variabilities is the cross correlation. When the statistic has strong peaks at some time lags the result is a good indication that the data have similarities. Then, we expect that there are correlation structures between the two signals (or that similar factors may influence both systems). On the other hand, when the statistic does not have strong peaks we will conclude that there is no similarity. Then, we expect that there is no correlation structure and that the dynamics of the systems have a different origin. However, this may not always be true because "no similarity" is not equivalent to "no correlation." That is to say, even when two signals are not similar, there are still possibilities that these systems have some kind of correlation structures (that is, these systems are interconnected or interrelated). To investigate this an approach from the viewpoint of a deterministic dynamical system is necessary. In this paper, we introduce such a method to investigate whether there are correlation structures in short-term variabilities among data, irrespective of whether data have similar or different long-term trends.

We first describe a current technique from the viewpoint of a deterministic dynamical system. Then we describe our technique. After describing these techniques, we will present our choice of discriminating statistics. Then, we will apply this algorithm to several cases using simulated time series data: (i) data have no trend, (ii) data have the same trends, and (iii) data have different trends. In each case, the data we use are both noise-free and contaminated by 10% Gaussian observational noise.

Based on the numerical experiments, we apply our method to real-world data. We select two specific systems of particular interest to us: human postural data (measured during quiet standing) of mediolateral and anteroposterior directions shown in Fig. 1(a), and electroencephalography (EEG), measured at C_Z of the unipolor 10–20 Jasper registration scheme [3], shown in Fig. 1(b). Both the data show irregular fluctuations and long-term trends. We show that there are correlation structures in the irregular fluctuations of human postural data, and there are correlation structures in the irregular fluctuations of EEG data depending on situations.

II. CURRENT TECHNOLOGY: THE RANDOM SHUFFLE SURROGATE METHOD

Surrogate methods have been proposed to investigate features of the data from the viewpoint of deterministic dynamical systems [4]. To investigate whether data can be fully described by independent and identically distributed (IID) random variables the random shuffle surrogate (RSS) method is useful [4]. We can apply the cross correlation to the origi-

*Electronic address: entomo@eie.polyu.edu.hk
†Electronic address: yoshito@sat.t.u-tokyo.ac.jp
‡Electronic address: ensmall@polyu.edu.hk

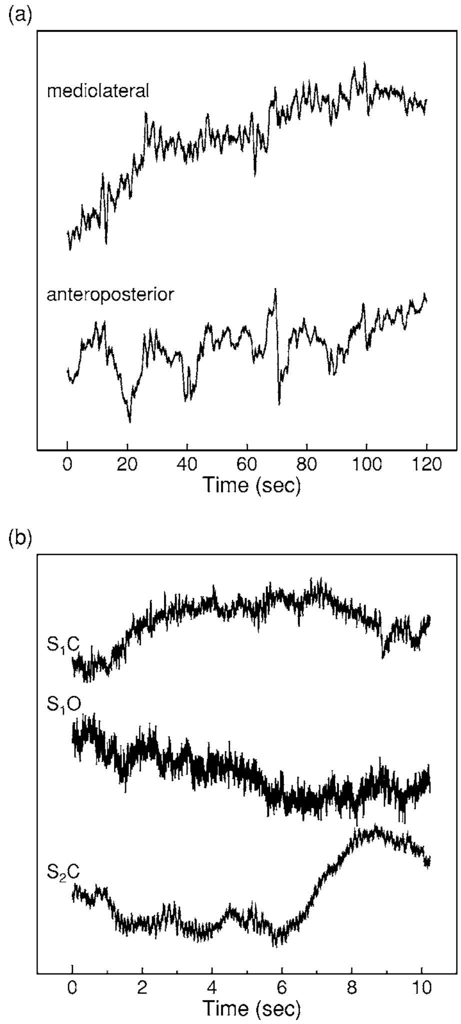

FIG. 1. Behavior of time series examined in this paper: Both (a) and (b) show irregular fluctuations and long-term trends. (a) is human postural data (measured during quiet standing) of mediolateral and anteroposterior directions. The data were measured at 100 Hz. The subject was healthy, stood barefoot, and had opened eyes. (b) is healthy human electroencephalography (EEG) data, where S_1 and S_2 are the subject identifiers and the third letter, O and C, identifies the condition: O=eyes open and resting, and C=eyes closed and resting. The measurement point is C_Z of the unipolor 10–20 Jasper registration scheme [3]. Measurements were not simultaneous and the data were digitized at 1024 Hz using a 12-bit A/D converter. The EEG impedances were less than 5 KΩ. The data were amplified, gain=18 000, and amplifier frequency cut-off settings of 0.03 Hz and 200 Hz were used. The ordinate axis in (a) and (b) is arbitrary.

nal and RSS data to investigate whether there are correlation structures. Although the RSS method is effective for time series with no trends (only irregular fluctuations), the algorithm is ineffective for data exhibiting slow trends because long-term trends are not preserved in RSS data. That is, irrespective of whether irregular fluctuations are modulated by long-term trends or not, the RSS method can indicate whether the *data* have some kind of dynamics, not only whether the *short-term variabilities* (irregular fluctuations) have some kind of dynamics [5].

Hence, when data exhibit short-term variabilities and long-term trends, to investigate whether there are correlation structures in short-term variabilities, we need to destroy local structures or correlations in short-term variabilities and preserve the global behaviors (trends). In the next section, we describe a method to generate data that can fulfill such conflicting conditions.

III. A DIFFERENT ALGORITHM: THE SMALL SHUFFLE SURROGATE METHOD

To produce surrogates that we can apply to test data even if it exhibits different long-term trends, the small shuffle surrogate (SSS) method has been proposed [6]. Furthermore, the method does not depend on the data distribution. The SSS method has proven to be effective for tackling data exhibiting short-term variabilities and long-term trends [5–7].

SSS data are generated as follows. Let the original data be $x(t)$, let $i(t)$ be the index of $x(t)$ [that is, $i(t)=t$, and so $x[i(t)]=x(t)$], let $g(t)$ be Gaussian random numbers (GRN), and $s(t)$ will be the surrogate data.

(i) Obtain $i'(t)=i(t)+Ag(t)$, where A is an amplitude (adding GRN to the index of the original data).

(ii) Sort $i'(t)$ by the rank order and let the index of $i'(t)$ be $\hat{i}(t)$ (rank order the perturbed index).

(iii) Obtain the surrogate data $s(t)=x[\hat{i}(t)]$ (reorder the original data with the perturbed index).

We have found that choosing $A=1.0$ is adequate for nearly all purposes, and the SSS data are very similar to the original data [5–7]. In the SSS data, local structures or correlations in irregular fluctuations (short-term variability) are destroyed and the global behaviors (trends) are preserved. Further details of the mechanism are provided in Refs. [5,7]. Then, the null hypothesis (NH) addressed by this algorithm is that irregular fluctuations (short-term variability) are independently distributed (ID) random variables (in other words, there is no short-term dynamics or determinism) [5–7]. Hence, when we apply the SSS method to multivariate data the NH is that there is no short-term correlation structure among data or that the irregular fluctuations are independent.

IV. HOW TO REJECT A NULL HYPOTHESIS

Discriminating statistics are necessary for hypothesis testing. After the calculation of the statistic, we need to inspect whether the NH should be rejected or not.

A. Discriminating statistics

Discriminating statistics are necessary for surrogate data hypothesis testing. The SSS method destroys local structures or correlations in irregular fluctuations (short-term variability) and preserves the global behaviors (trends). That is, the SSS data change the flow of information in the irregular fluctuations. It is preferable to use discriminating statistics that can reflect features of the surrogate method. Hence, we choose to use the cross correlation function (CC) and the average mutual information (AMI) as discriminating statistics. The CC—an estimate of the linear correlation between two signals—and the AMI—a nonlinear version of the CC—can determine, on average, how much one learns between two signals [8].

We note here that it is widely observed that estimating AMI is difficult [9]. The major reason is that it is not easy to estimate the underlying probability distribution reliably. To reduce this problem a new method has been proposed where an adaptive partition is applied [10]. However, the SSS data have the same probability distribution (rank distribution) as the original data. In this case, we consider that the influence due to using different data (the original data and the SSS data) for estimating the joint probability distribution is not large, and we find that there is not significant bias between the estimated joint probability distribution of the original data and the SSS data. Hence, we expect that it is relatively straightforward to compare the AMI of the original data and the SSS data.

B. Monte Carlo hypothesis testing

After the calculation of these statistics, we need to inspect whether a NH shall be rejected or not. If there is sufficient difference between the original data and surrogate data, the NH is rejected. In this case, we consider that the original and the surrogate data would not come from the same population. If there is no significant difference, one may not reject the NH. In this case, we consider that the original and the surrogate data may come from the same population.

For this inspection, we employ Monte Carlo hypothesis testing and inspect whether the estimated statistics of the original data fall within or outside the statistical distribution of the surrogate data [11]. When the statistics fall within the distribution of the surrogate data, we conclude that the hypothesis may not be rejected. In this paper, we generate 99 SSS data and hence the significance level is between 0.01 and 0.02 for a one-sided test with two nonindependent statistics [12].

Although the multiple-comparison problem is common in surrogate data applications, we use two discriminating statistics, the CC and the AMI, as complementary statistics. This is because we have found that a statistic does not work but the other works well in some test systems [5]. One of the systems is the logistic map [13]. The logistic map is given as $x(t+1)=4.0x(t)[1-x(t)]$; we use 5000 data points and the data is noise-free. We apply the SSS method to the data, and we use the autocorrelation (AC) and the AMI as discriminating statistics. While the logistic map has clear nonlinear dynamics, the AC of the original data falls within the distribution of the SSS data, and the AMI of the original data falls outside the distribution of the SSS data. This result indicates that only one statistic is insufficient for some cases. Hence, to avoid this problem we adopt two discriminating statistics, the CC and the AMI, for our tests. Furthermore, as shown later, we note that when irregular fluctuations are independent or have no correlation structure, *both* the CC and the AMI of the data must fall within the distribution of the SSS data.

Also, we show plots of both the CC and the AMI as a function of time lag (in other words, the variation of the CC and the AMI with lag is shown). However, in all cases the hypothesis testing is robustly conducted for a fixed small value of lag (for example, lag=1 or −1). In fact, we expect that it is only a meaningful test statistic for small lag, because the CC and the AMI of the original and surrogate data will coincide for large lag. The plots of the CC and the AMI as a function of lag are provided for information only.

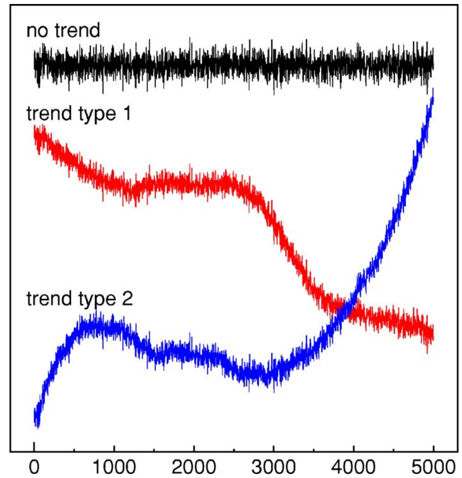

FIG. 2. (Color online) Behavior of our numerical data, irregular fluctuations with no trend, those with trends, and those with other trends. In this panel, the irregular fluctuations are Gaussian random numbers and both the long-term trends are artificial.

V. NUMERICAL EXAMPLES

We now demonstrate the application of our algorithm to various simulated time series data, and confirm our theoretical arguments with several examples. Broadly speaking, we use two types of time series, data with no trend and data with long-term trends. Furthermore, we use two types of long-term trends, the same trends and different trends (see Fig. 2). In all cases, we use $A=1.0$, for generating SSS data, generate 99 SSS data, the number of data points is 5000, and the data is both noise-free and subsequently contaminated by 10% (20 dB) Gaussian observational noise.

A. Data with no long-term trend

The first application is to two scalar time series with no long-term trend. To study irregular fluctuations that have correlation structures (interconnected or interrelated dynamics), we use the following models.

(i) The coupled linear autoregressive moving average (ARMA) model given by

$$x(t) = 24.0 + 0.58x(t-1) - 0.22x(t-2) - 0.5y(t-1) + 0.13y(t-5) + \eta(t),$$

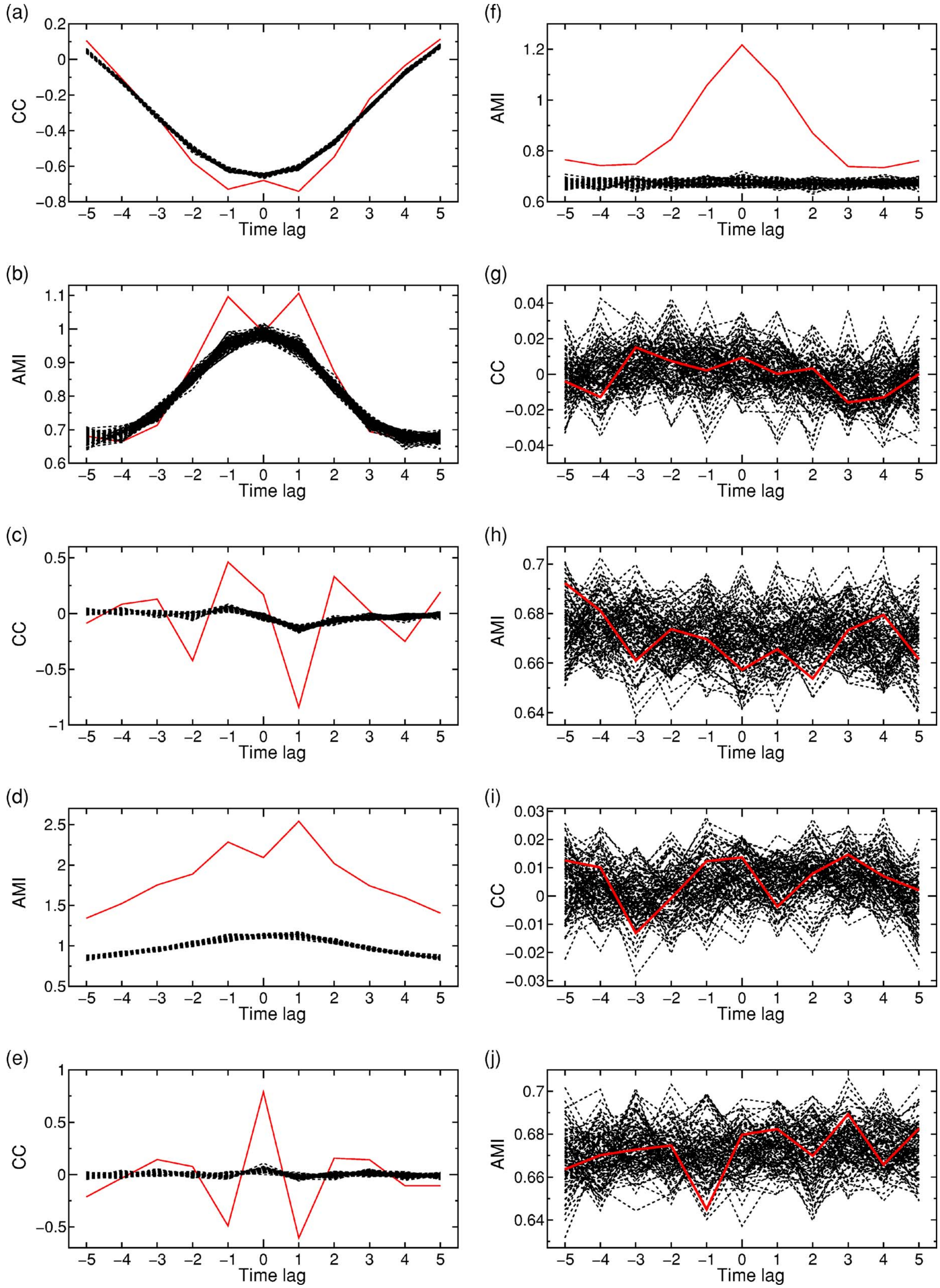

FIG. 3. (Color online) A plot of (a), (c), (e), (g), and (i) cross correlation (CC), and (b), (d), (f), (h), and (j) the average mutual information (AMI), where data have no trend and we use 99 SSS data. (a) and (b) are the coupled linear ARMA model, (c) and (d) are the Ikeda map, (e) and (f) are a chaotic neural network (CNN), (g) and (h) are Gaussian random numbers (GRN), and (i) and (j) are the x component of the coupled linear ARMA model and that of the Ikeda map. The solid line is the original data and the dotted lines are the SSS data.

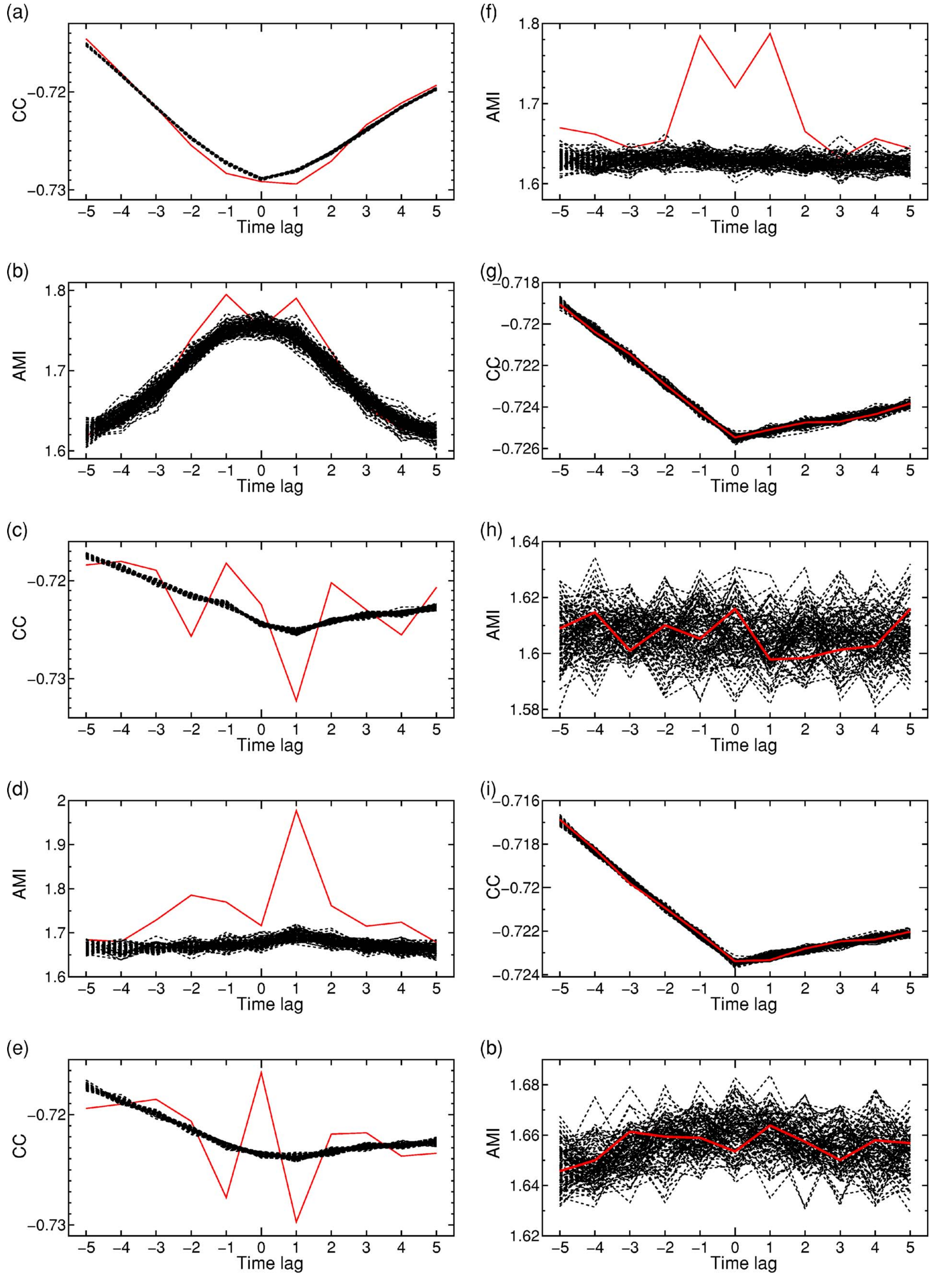

FIG. 4. (Color online) A plot of CC and AMI, where data have different long-term trends and we use 99 SSS data. The notation is the same as in Fig. 3.

$$y(t) = 29.0 - 0.75x(t-1) + 0.26x(t-3) + 0.5y(t-1) - 0.25y(t-2) + \xi(t),$$

where $\eta(t)$ and $\xi(t)$ are Gaussian dynamic noise with standard deviation 0.07 [14].

(ii) The Ikeda map given by

$$f(x,y) = [1 + \mu(x \cos \theta - y \sin \theta), \mu(x \sin \theta + y \cos \theta)],$$

where $\theta = a - b/(1 + x^2 + y^2)$ with $\mu = 0.83$, $a = 0.4$, and $b = 6.0$ [15].

In all cases, we use $x(t)$ and $y(t)$ as the observational data.

We also use a more complex model. Many measured physical quantities can be seen as an average derived from subsystems or microsystems. However, in the scalar time series analysis it is assumed to be a single probe of the system that is investigated. Also, the ensemble operation can cancel individuality and make the collective behavior stochastic. One of the examples is EEG. To investigate such a more complex and practical case, we use a chaotic neural network (CNN) [16]. The CNN system is given by

$$y_i(t+1) = ky_i(t) + \sum_{j=1}^{n} w_{ij} z_j(t) - \alpha z_i(t) + a,$$

$$z_i(t+1) = f[y_i(t+1)],$$

where n is the number of neurons, $z_i(t)$ and $y_i(t)$ are the output and the internal states, f is the logistic function $f(y) = 1/[1 + \exp(-y/\epsilon)]$, k and α are the control parameters, and w_{ij} is the synaptic weight from the jth neuron to ith neuron [16]. We use 20 neurons to compose the network [17]. The ensemble mean value $x(t)$ of N neuron is defined as $x(t) = \frac{1}{N}\sum_{i=1}^{N} y_i(t)$, which can be regarded as a simple model of EEG data. We use two ensemble mean values as the observational data $x_1(t) = \frac{1}{10}\sum_{i=1}^{10} y_i(t)$ and $x_2(t) = \frac{1}{10}\sum_{i=11}^{20} y_i(t)$, where the ten neurons used in one ensemble mean are not used in another ensemble mean.

As an example of irregular fluctuations where the data have dynamics but they are independent, we use data generated by the x component of the linear ARMA model and the Ikeda map as mentioned previously. Furthermore, to investigate when data have no dynamics and are independent, we use GRN as irregular fluctuations.

Figure 3 shows the results. When data are GRN or data are x components of the coupled linear ARMA model and that of the Ikeda map, both of the CC and AMI of the original data fall within the distributions of the SSS data. However, in other cases (when there are some kind of correlation structures between two signals), both of the CC and AMI are distinct. These results indicate that we can discriminate correctly whether there are correlation structures between two signals. Here, we note that some differences clearly appear when the time lag is relatively small, because the information in the systems is not retained for longer periods of time. When the data are contaminated by 10% observational noise, the results obtained are essentially the same.

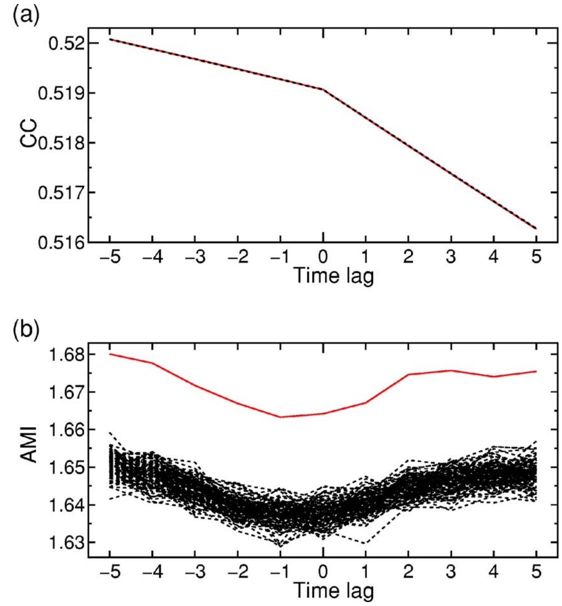

FIG. 5. (Color online) A plot of the CC and AMI for the postural data: (a) CC and (b) AMI. The solid line is the original data and the dotted lines are the SSS data.

B. Data with-long term trends

To investigate irregular fluctuations with long-term trends, irregular fluctuations generated using the same models as above are added to the artificial trends, where we use two different trends, trend type 1 and type 2 shown in Fig. 2. Then, we can investigate when the long-term trends are the same and different. The level of additional data to the data is equivalent to 10% (20 dB) observational noise at each case. Figure 2 shows the behaviors. We show results when irregular fluctuations have different trends.

Figure 4 shows the results. When irregular fluctuations are Gaussian random numbers or data are x component of the coupled linear ARMA model and that of the Ikeda map, both of the CC and the AMI of the original data fall within the distributions of the SSS data. However, in other cases (when there are some kind of correlation structures between two signals), both of the CC and AMI are distinct. These results indicate that we can discriminate correctly whether there are correlation structures between two even when data have long-term trends. In all cases, especially when the time lag is larger, behaviors of the CC and the AMI of the SSS data are very similar to that of the original data. This indicates that the local structures are destroyed and the global structures are preserved in the SSS data. When data have the same trends and data are contaminated by 10% observational noise, the results are essentially the same.

These results indicate that we can discriminate correctly whether there are correlation structures between two signals, irrespective of whether data have the same or different long-term trends. Therefore, we conclude that applying the SSS method can detect whether there are correlation structures or not using the CC and AMI.

VI. APPLICATIONS

We now present the application of our proposed method to two experimental systems: (i) human postural data (measured during "quiet standing") of mediolateral and anteroposterior directions shown in Fig. 1(a), and (ii) EEG data measured at different positions shown in Fig. 1(b). Both the data seem to have trends. We use 12 000 data points (two minutes) for the postural data, and 10 240 data points (10 s) for the EEG data. In all cases we use $A=1.0$ and generate 99 SSS data.

Figure 5 shows the result of applying the SSS method to the postural data. The figure shows that although the CC of the original data falls within the distributions of SSS data, the AMI falls outside of the distribution. Hence, we consider that the irregular fluctuations of mediolateral and anteroposterior data have correlation structures. This result is in agreement with our understanding because we consider that to keep standing the posture balance of mediolateral and anteroposterior directions is controlled and the interrelation is necessary.

Figure 6 shows the result of applying the SSS method to the EEG data. Figures 6(a) and 6(b) show that both the CC and the AMI of the original data fall within the distributions of SSS data. This result indicates that the irregular fluctuations of the same subject but under different conditions do not have correlation structures. Figures 6(c) and 6(d) show that both the CC and the AMI of the original data fall outside the distributions of SSS data, where the difference of the AMI between the original and SSS data is distinct, although that of the CC is only slightly. As the subjects of the data are different, these data are not interconnected or interrelated clearly. The data are measured under the same condition, eyes closed and resting (these conditions reflect a relaxed state). As the brains are expected to be idle, similar activity is expected to occur in both cases. As a result, we consider that the irregular fluctuations of the data become similar [18].

It should be noted that Figs. 5 and 6 do not show strong peaks as typified by Figs. 3 and 4. In these cases we cannot easily decide whether there are correlation structures between the two signals using the CC or the AMI of only the original data. Hence, we conclude that our method was effective to investigate the relation of these data.

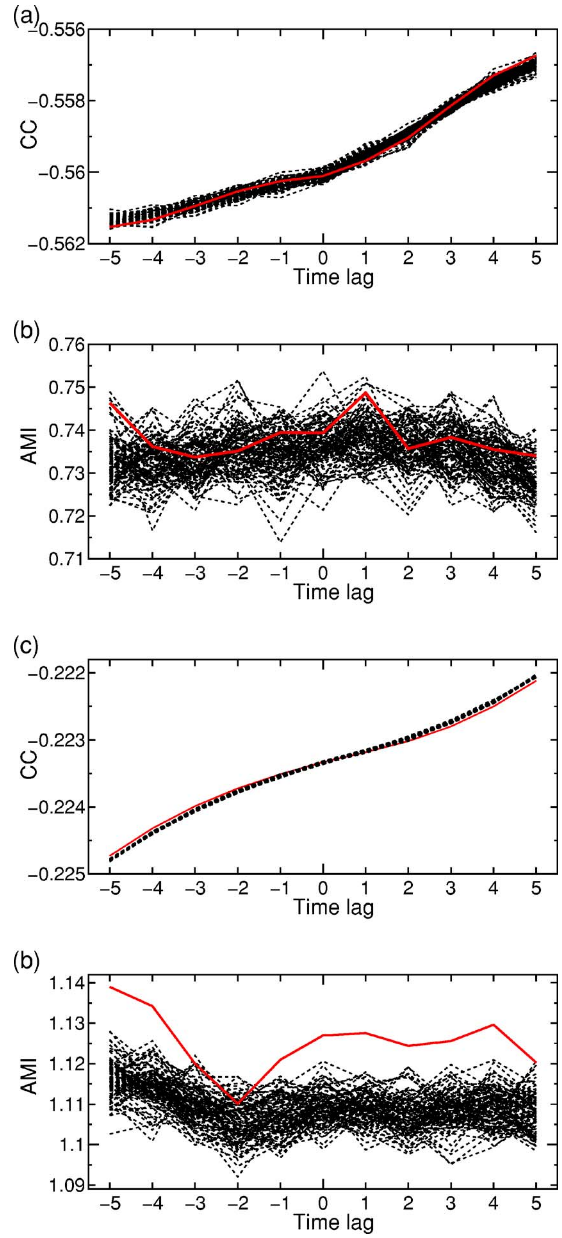

FIG. 6. (Color online) A plot of (a) and (c) CC, and (b) and (d) AMI for the EEG data: (a) and (b) S_1C and S_1O (the same subject and different conditions), and (c) and (d) S_1C and S_2C (different subjects and the same condition). For the explanation of this notation, see Fig. 1. The solid line is the original data and the dotted lines are the SSS data.

VII. CONCLUSION

We have described an algorithm for investigating whether there are correlation structures in irregular fluctuations of multivariate time series, even if they exhibit long-term trends. We have demonstrated the application of this algorithm to computational examples using the CC and the AMI as discriminating statistics. Our arguments and computational examples show that this algorithm succeeds in testing correlation structures in irregular fluctuations irrespective of whether the data have long-term trends.

ACKNOWLEDGMENTS

The authors would like to acknowledge Professor T. Nomura (Osaka University) and Professor P. E. Rapp (Drexel University) for providing us with the postural data and the EEG data, respectively. This work was supported by a direct allocation grant from Hong Kong Polytechnic University (Grant No. A-PG58).

[1] The term "correlation structures" we use means any structures, irrespective of whether the structures are linear or nonlinear.
[2] The name of the game is that one event sets off a chain of diverse events, and the event influences another event one after another. As a result, something astonishing has happened.
[3] H. Jasper, Electroencephalogr. Clin. Neurophysiol. **10**, 371 (1958).
[4] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, Physica D **58**, 77 (1992).
[5] T. Nakamura and M. Small, Int. J. Bifurcation Chaos Appl. Sci. Eng. (to be published).
[6] T. Nakamura and M. Small, Phys. Rev. E **72**, 056216 (2005).
[7] T. Nakamura and M. Small, Physica A **366**, 377 (2006).
[8] H. D. I. Abarbanel, *Analysis of Observed Chaotic Data* (Springer-Verlag, New York, 1996).
[9] C. J. Cellucci, A. M. Albano, and P. E. Rapp, Phys. Rev. E **67**, 066210 (2003).
[10] C. J. Cellucci, A. M. Albano, and P. E. Rapp, Phys. Rev. E **71**, 066208 (2005).
[11] J. Theiler and D. Prichard, Physica D **94**, 221 (1996).
[12] The significance level of each test is 0.01. If two statistics are identical (dependent), the significance level for the proposed test is 0.01. If the statistics are independent, the significance level for the test is given by $1.0-0.99\times 0.99=0.0199$. Hence, the reality should be somewhere in-between.
[13] R. M. May, Nature (London) **261**, 459 (1976).
[14] To obtain this ARMA model we applied a reduced autoregressive model procedure [19] to a human postural data during quiet standing of left and right directions. The data was measured at 20 Hz. The subject was healthy, stood barefoot, and had opened eyes.
[15] K. Ikeda, Opt. Commun. **30**, 257 (1979).
[16] K. Aihara, T. Takabe, and M. Toyoda, Phys. Lett. A **144**, 333 (1990).
[17] We use $a=0.35$, $k=0.7$, $\alpha=1.0$, and $\epsilon=0.025$. The largest Lyapunov exponent estimated directly from the model (not an embedding of the data) is 0.361 and the Lyapunov dimension is 7.655. This system is therefore consistent with chaotic dynamics and a deterministic fractal attractor.
[18] Discussions on how to understand EEG data apply only to the specific EEG data used in this paper. We do not wish to overinterpret how the brain works.
[19] T. Nakamura, K. Judd, and A. I. Mees, Int. J. Bifurcation Chaos Appl. Sci. Eng. **13**, 1263 (2003).