Radial basis function approach to nonlinear Granger causality of time series

Nicola Ancona,[1] Daniele Marinazzo,[2,3,4] and Sebastiano Stramaglia[2,3,4]
[1]*Istituto di Studi sui Sistemi Intelligenti per l'Automazione, CNR, Bari, Italy*
[2]*TIRES-Center of Innovative Technologies for Signal Detection and Processing, Università di Bari, Bari, Italy*
[3]*Dipartimento Interateneo di Fisica, Bari, Italy*
[4]*Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Bari, Italy*
(Received 3 May 2004; revised manuscript received 1 July 2004; published 23 November 2004)

We consider an extension of Granger causality to nonlinear bivariate time series. In this frame, if the prediction error of the first time series is reduced by including measurements from the second time series, then the second time series is said to have a causal influence on the first one. Not all the nonlinear prediction schemes are suitable to evaluate causality; indeed, not all of them allow one to quantify how much knowledge of the other time series counts to improve prediction error. We present an approach with bivariate time series modeled by a generalization of radial basis functions and show its application to a pair of unidirectionally coupled chaotic maps and to physiological examples.

DOI: 10.1103/PhysRevE.70.056221 PACS number(s): 05.10.−a, 87.10.+e, 89.70.+c

I. INTRODUCTION

Identifying causal relations among simultaneously acquired signals is an important problem in computational time series analysis and has applications in economy [1,2], EEG analysis [3], human cardiorespiratory system [4], interaction between heart rate and systolic arterial pressure [5], and many others. Several papers dealt with this problem, relating it to identification of interdependence in nonlinear dynamical systems [6,7] or to estimates of information rates [8,9]. Some approaches modeled data by oscillators and concentrated on the phases of the signals [10]. One major approach to analyze the causality between two time series is to examine if the prediction of one series could be improved by incorporating information of the other, as proposed by Granger [1] in the context of linear regression models of stochastic processes. In particular, if the prediction error of the first time series is reduced by including measurements from the second time series in the linear regression model, then the second time series is said to have a causal influence on the first time series. By exchanging the roles of the two time series, one can address the question of the causal influence in the opposite direction. It is worth stressing that, within this definition of causality, flow of time plays a major role in making inference, from time series data, depending on direction. Since Granger causality was formulated for linear models, its application to nonlinear systems may not be appropriate. In this paper we consider the problem of extending Granger causality definition to nonlinear problems.

In the next section we review the original approach by Granger while describing our point of view about its nonlinear extension; we also propose a method which fulfills the requirements a prediction scheme should satisfy to analyze causality. Our method exploits radial basis functions, an algorithm initially proposed to perform exact interpolation of a set of data points in a multidimensional space (see, e.g., [11]). In Sec. III we show application of the proposed method to simulated and real examples. Some conclusions are drawn in Sec. IV.

II. GRANGER CAUSALITY

A. Linear modeling of bivariate time series

We briefly recall the vector autoregressive (VAR) model which is used to define linear Granger causality [1]. Let $\{x_i\}_{i=1,\ldots,N}$ and $\{y_i\}_{i=1,\ldots,N}$ be two time series of N simultaneously measured quantities. In the following we will assume that time series are stationary. For $k=1$ to M (where $M=N-m$, m being the order of the model), we denote $x^k=x_{k+m}$, $y^k=y_{k+m}$, $\mathbf{X}^k=(x_{k+m-1},x_{k+m-2},\ldots,x_k)$, and $\mathbf{Y}^k=(y_{k+m-1},y_{k+m-2},\ldots,y_k)$ and we treat these quantities as M realizations of the stochastic variables (x, y, $\mathbf{X}$, $\mathbf{Y}$). The following model is then considered [12]:

$$x=\mathbf{W}_{11}\cdot\mathbf{X}+\mathbf{W}_{12}\cdot\mathbf{Y},$$

$$y=\mathbf{W}_{21}\cdot\mathbf{X}+\mathbf{W}_{22}\cdot\mathbf{Y}, \tag{1}$$

$\{\mathbf{W}\}$ being four m-dimensional real vectors to be estimated from data by standard least-squares techniques. Let us call ϵ_{xy} and ϵ_{yx} the prediction errors of this model, defined as the estimated variances of $x-\mathbf{W}_{11}\cdot\mathbf{X}-\mathbf{W}_{12}\cdot\mathbf{Y}$ and $y-\mathbf{W}_{21}\cdot\mathbf{X}-\mathbf{W}_{22}\cdot\mathbf{Y}$, respectively. We also consider autoregressive (AR) predictions of the two time series—i.e., the model

$$x=\mathbf{V}_1\cdot\mathbf{X},$$

$$y=\mathbf{V}_2\cdot\mathbf{Y}, \tag{2}$$

$\mathbf{V}_1$ and $\mathbf{V}_2$ to be estimated by least squares fit. The estimate of the variance of $x-\mathbf{V}_1\cdot\mathbf{X}$ is called ϵ_x (the prediction error when x is predicted solely on the basis of knowledge of its past values); similarly, ϵ_y is the variance of $y-\mathbf{V}_2\cdot\mathbf{Y}$. If the prediction of x improves by incorporating the past values of $\{y_i\}$—i.e., ϵ_{xy} is smaller than ϵ_x—then y has a causal influence on x. Analogously, if ϵ_{yx} is smaller than ϵ_y, then x has a causal influence on y. Calling $c_1=\epsilon_x-\epsilon_{xy}$ and $c_2=\epsilon_y-\epsilon_{yx}$, a directionality index can be introduced:

$$D = \frac{c_2 - c_1}{c_1 + c_2}. \tag{3}$$

The index D varies from 1 in the case of unidirectional influence ($x \to y$) to -1 in the opposite case ($y \to x$), with intermediate values corresponding to bidirectional influence. According to this definition of causality, the following property holds for M sufficiently large: *if* $\mathbf{Y}$ *is uncorrelated with* $\mathbf{X}$ *and* x, *then* $\epsilon_x = \epsilon_{xy}$. This means that in this case VAR and AR modelings of the $\{x_i\}$ time series coincide. Analogously *if* $\mathbf{X}$ *is uncorrelated with* $\mathbf{Y}$ *and* y, *then* $\epsilon_y = \epsilon_{yx}$. It is clear that these properties are fundamental and make the linear prediction approach suitable to evaluate causality. On the other hand, for nonlinear systems higher-order correlations may be relevant. Therefore, we propose that any prediction scheme providing a nonlinear extension of Granger causality should satisfy the following property: (P1) *if* $\mathbf{Y}$ *is statistically independent of* $\mathbf{X}$ *and* x, *then* $\epsilon_x = \epsilon_{xy}$; *if* $\mathbf{X}$ *is statistically independent of* $\mathbf{Y}$ *and* y, *then* $\epsilon_y = \epsilon_{yx}$. In a recent paper [13], use of a locally linear prediction scheme [14] has been proposed to evaluate nonlinear causality. In this scheme, the joint dynamics of the two time series is reconstructed by delay vectors embedded in an Euclidean space; in the delay embedding space a locally linear model is fitted to data. The approach described in [13] satisfies property P1 only if the number of points in the neighborhood of each reference point, where linear fit is done, is sufficiently high to establish good statistics; however, linearization is valid only for small neighborhoods. It follows that this approach to nonlinear causality requires very long time series to satisfy P1. In order to construct methods working effectively with moderately long time series, in the next subsection we will characterize the problem of extending Granger causality as the one of finding classes of nonlinear models satisfying property P1.

B. Nonlinear models

What is the most general class of nonlinear models which satisfy P1? The complete answer to this question is matter for further study. Here we only give a partial answer—i.e., the following family of models:

$$x = \mathbf{w}_{11} \cdot \mathbf{\Phi}(\mathbf{X}) + \mathbf{w}_{12} \cdot \mathbf{\Psi}(\mathbf{Y}),$$

$$y = \mathbf{w}_{21} \cdot \mathbf{\Phi}(\mathbf{X}) + \mathbf{w}_{22} \cdot \mathbf{\Psi}(\mathbf{Y}), \tag{4}$$

where $\{\mathbf{w}\}$ are four n-dimensional real vectors, $\mathbf{\Phi} = (\varphi_1, \ldots, \varphi_n)$ are n given nonlinear real functions of m variables, and $\mathbf{\Psi} = (\psi_1, \ldots, \psi_n)$ are n other real functions of m variables. Given $\mathbf{\Phi}$ and $\mathbf{\Psi}$, model (4) is a linear function in the space of features φ and ψ; it depends on $4n$ variables, the vectors $\{\mathbf{w}\}$, which must be fixed to minimize the prediction errors

$$\epsilon_{xy} = \frac{1}{M}\sum_{k=1}^{M} [x^k - \mathbf{w}_{11} \cdot \mathbf{\Phi}(\mathbf{X}^k) - \mathbf{w}_{12} \cdot \mathbf{\Psi}(\mathbf{Y}^k)]^2;$$

$$\epsilon_{yx} = \frac{1}{M}\sum_{k=1}^{M} [y^k - \mathbf{w}_{21} \cdot \mathbf{\Phi}(\mathbf{X}^k) - \mathbf{w}_{22} \cdot \mathbf{\Psi}(\mathbf{Y}^k)]^2. \tag{5}$$

We also consider the model

$$x = \mathbf{v}_1 \cdot \mathbf{\Phi}(\mathbf{X}),$$

$$y = \mathbf{v}_2 \cdot \mathbf{\Psi}(\mathbf{Y}), \tag{6}$$

and the corresponding prediction errors ϵ_x and ϵ_y.

Now we prove that model (4) satisfies P1. Let us suppose that $\mathbf{Y}$ is statistically independent of $\mathbf{X}$ and x. Then, for each $\mu = 1, \ldots, n$ and for each $\lambda = 1, \ldots, n$, $\psi_\mu(\mathbf{Y})$ is uncorrelated with x and with $\varphi_\lambda(\mathbf{X})$. It follows that

$$\begin{aligned} \epsilon_{xy} &= \mathrm{var}[x - \mathbf{w}_{11} \cdot \mathbf{\Phi}(\mathbf{X}) - \mathbf{w}_{12} \cdot \mathbf{\Psi}(\mathbf{Y})] \\ &= \mathrm{var}[x - \mathbf{w}_{11} \cdot \mathbf{\Phi}(\mathbf{X})] + \mathrm{var}[\mathbf{w}_{12} \cdot \mathbf{\Psi}(\mathbf{Y})]. \end{aligned} \tag{7}$$

The vectors $\{\mathbf{w}\}$ must be fixed to minimize the prediction error ϵ_{xy}: from the equation above it follows that, for large M, the minimum corresponds to $\mathbf{w}_{12} = 0$; hence, models (4) and (6) of the $\{x_i\}$ time series coincide. The same argument may be used exchanging x and y. This proves that P1 holds.

C. Radial basis functions

In this subsection we propose a strategy to choose the functions $\mathbf{\Phi}$ and $\mathbf{\Psi}$, in model (4), in the frame of radial basis functions (RBF) methods. Fixed $n \ll M$, n centers $\{\tilde{\mathbf{X}}^\rho\}_{\rho=1}^n$, in the space of $\mathbf{X}$ vectors, are determined by a clustering procedure applied to data $\{\mathbf{X}^k\}_{k=1}^M$. Analogously n centers $\{\tilde{\mathbf{Y}}^\rho\}_{\rho=1}^n$, in the space of $\mathbf{Y}$ vectors, are determined by a clustering procedure applied to data $\{\mathbf{Y}^k\}_{k=1}^M$. We then make the following choice:

$$\varphi_\rho(\mathbf{X}) = \exp(-\|\mathbf{X} - \tilde{\mathbf{X}}^\rho\|^2/2\sigma^2), \quad \rho = 1, \ldots, n,$$

$$\psi_\rho(\mathbf{Y}) = \exp(-\|\mathbf{Y} - \tilde{\mathbf{Y}}^\rho\|^2/2\sigma^2), \quad \rho = 1, \ldots, n, \tag{8}$$

σ being a fixed parameter, whose order of magnitude is the average spacing between the centers. The centers $\{\tilde{\mathbf{X}}^\rho\}$ are prototypes of the $\mathbf{X}$ variables; hence, φ functions measure the similarity to these typical patterns. Analogously, ψ functions measure the similarity to typical patterns of $\mathbf{Y}$. Many clustering algorithms may be applied to find prototypes; for example, in our experiments we use fuzzy c means [15].

Some remarks are in order. First, we observe that the models described above may trivially be adapted to handle the case of reconstruction embedding of the two time series in a delay coordinate space, as described in [13]. Second, we stress that in Eqs. (4) x and y are modeled as the sum of two contributions, one depending solely on $\mathbf{X}$ and the other dependent on $\mathbf{Y}$. Obviously better prediction models for x and y exist, but they would not be useful to evaluate causality unless they would satisfy P1. This requirement poses a limit to the level of detail at which the two time series may be described, if one is looking at causality relationships. The justification of the model we propose here, based on regularization theory [16], is sketched in the Appendix. In the Appendix we also recall the standard RBF modeling of the bivariate time series.

D. Empirical risk and generalization error

In the previous subsections the prediction error has been identified as the empirical risk, although there is a difference between these two quantities as statistical learning theory (SLT) [17] shows. The deep connection between empirical risk and generalization error deserves a comment here. First of all we want to point out that the ultimate goal of a predictor and in general of any supervised machine $x=f(\mathbf{X})$ [18] is *to generalize*—that is, to correctly predict the output values x corresponding to never seen before input patterns $\mathbf{X}$ (for definiteness we consider the case of predicting x on the basis of the knowledge of $\mathbf{X}$). A measure of the generalization error of such a machine f is the *risk* $R[f]$ defined as the expected value of the loss function $V(x,f(\mathbf{X}))$:

$$R[f]=\int dx\, d\mathbf{X}\, V(x,f(\mathbf{X}))P(x,\mathbf{X}), \tag{9}$$

where $P(x,\mathbf{X})$ is the probability density function underlying the data. A typical example of loss function is $V(x,f(\mathbf{X}))=(x-f(\mathbf{X}))^2$ and in this case the function minimizing $R[f]$ is called the *regression function*. In general P is unknown and so we cannot minimize the risk. The only data we have are M observations (examples) $S=\{(x^k,\mathbf{X}^k)\}_{k=1}^M$ of the random variables x and $\mathbf{X}$ drawn according to $P(x,\mathbf{X})$. Statistical learning theory [17] as well as regularization theory [16] provides upper bounds of the generalization error of a learning machine f. Inequalities of the following type may be proven:

$$R[f]\leqslant \epsilon_x+\mathcal{C}, \tag{10}$$

where

$$\epsilon_x=\frac{1}{M}\sum_{k=1}^{M}[x^k-f(\mathbf{X}^k)]^2 \tag{11}$$

is the *empirical risk*, which measures the error on the training data. $\mathcal{C}$ is a measure of the *complexity* of machine f and it is related to the so-called Vapnik-Chervonenkis (VC) dimension. Predictors with low complexity guarantee low generalization error because they avoid overfitting. When the complexity of the functional space where our predictor "lives" is *small*, then the empirical risk is a good approximation of the generalization error. The models we deal with in this work verify such constraint. In fact, linear predictors have a finite VC dimension, equal to the size of the space where the input patterns live, and predictors expressed as linear combinations of radial basis functions are smooth. In conclusion empirical risk is a good measure of the generalization error for the predictors we are considering here and so it can be used to construct measures of causality between time series [19].

III. EXPERIMENTS

In order to demonstrate the usefulness of the proposed approach, in this section we study two examples: a pair of unidirectionally coupled chaotic maps and two physiological problems.

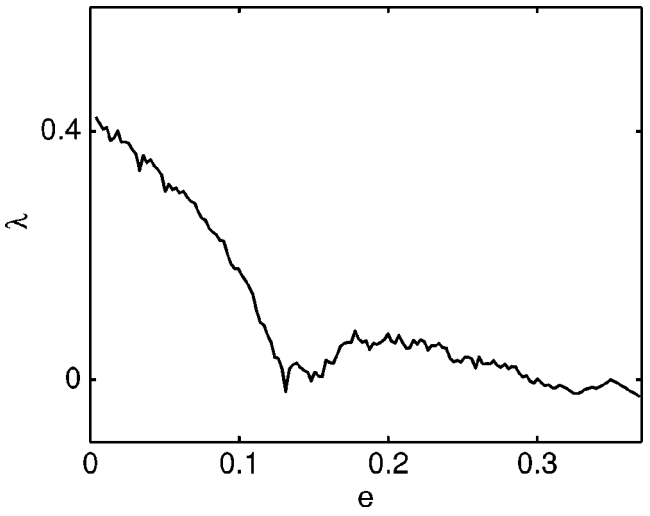

FIG. 1. The second Lyapunov exponent of the coupled maps (12) is plotted versus coupling e.

A. Chaotic maps

Let us consider the following pair of noisy logistic maps:

$$x_{n+1}=a\,x_n\,(1-x_n)+s\,\eta_{n+1},$$

$$y_{n+1}=(1-e)\,a\,y_n\,(1-y_n)+e\,a\,x_n\,(1-x_n)+s\xi_{n+1}; \tag{12}$$

$\{\eta\}$ and $\{\xi\}$ are unit variance Gaussianly distributed noise terms; the positive parameter s determines their relevance. Using $s\leqslant 0.07$, the time series is not observed to diverge. We fix $a=3.8$, and $e\in[0,1]$ represents the coupling $x\rightarrow y$. In the noise-free case ($s=0$), a transition to synchronization ($xn=yn$) occurs at $e=0.37$. We evaluate the Lyapunov exponents by the method described in [20]: the first exponent is 0.43, and the second exponent depends on e and is depicted in Fig. 1 for $e<0.37$ (it becomes negative for $e>0.37$). For several values of e, we have considered runs of 10^5 iterations, after 10^5 transients, and evaluated the prediction errors by Eqs. (4) and (6), with $m=1$, $n=100$, and $\sigma=0.05$. In Fig. 2(a) we depict, in the noise-free case, the curves representing c_1 and c_2 versus coupling e. In Figs. 2(b)–2(d) we depict the directionality index D versus e, in the noise-free case and for $s=0.01$ and $s=0.07$, respectively. In the noise-free case we find $D=1$; i.e., our method revealed unidirectional influence. As the noise increases, also the minimum value of e, which renders unidirectional coupling detectable, increases.

B. Physiological data

As a real example, we consider time series of heart rate and breath rate of a sleeping human suffering from sleep apnea (10 min from data set B of the Santa Fe Institute time series contest held in 1991, available in the Physionet data bank [21]). There is growing evidence that suggests a causal link between sleep apnea and cardiovascular disease [22], although the exact mechanisms that underlie this relationship remain unresolved [23]. Figure 3 clearly shows that bursts of

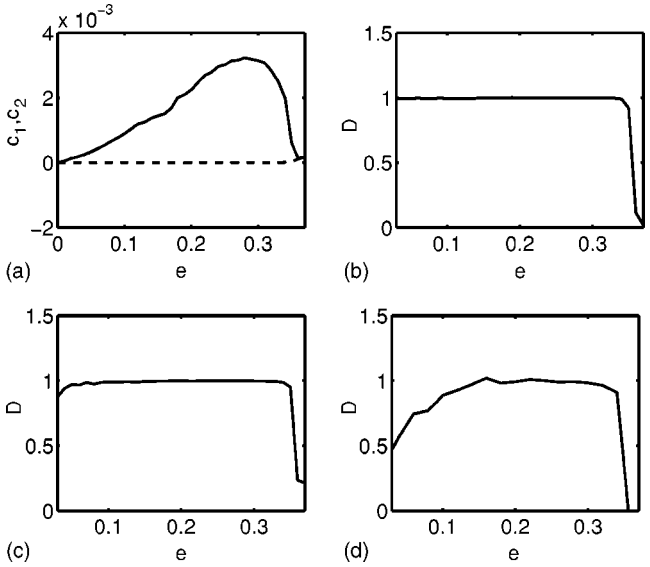

FIG. 2. (a) For the noise-free case of coupled maps (12), $c_1 = \epsilon_x - \epsilon_{xy}$ (dashed line) and $c_2 = \epsilon_y - \epsilon_{yx}$ (solid line) are plotted versus coupling e. (b) The directionality index D (see the text) is plotted versus e in the noise-free case. (c) The directionality index D is plotted versus e, $s=0.01$. (d) D is plotted versus e, $s=0.07$.

the patient breath and cyclical fluctuations of heart rate are interdependent. We fix $n=50$, $\sigma=0.5$ and vary m in $\{1,2,\ldots,20\}$. In Fig. 4 we depict ϵ_x (x representing heart rate) and ϵ_y (y representing breath) as a function of m. The value of m, providing the best model of time series, corresponds to the knee of these curves: a greater value would result in a more complicated model without a significant improvement of the prediction error. In Fig. 5 we depict the quantities $\delta_1 = c_1/\epsilon_x$ and $\delta_2 = c_2/\epsilon_y$, which measure the influence of one variable on the other. Since the curve δ_2 is always above δ_1, we may conclude that the causal influence of heart rate on breath is stronger than the reverse [24]. Concerning the directional index D, we evaluate it at the peaks of δ curves—i.e., at $m=5$—and obtain $D=0.76$, a positive and

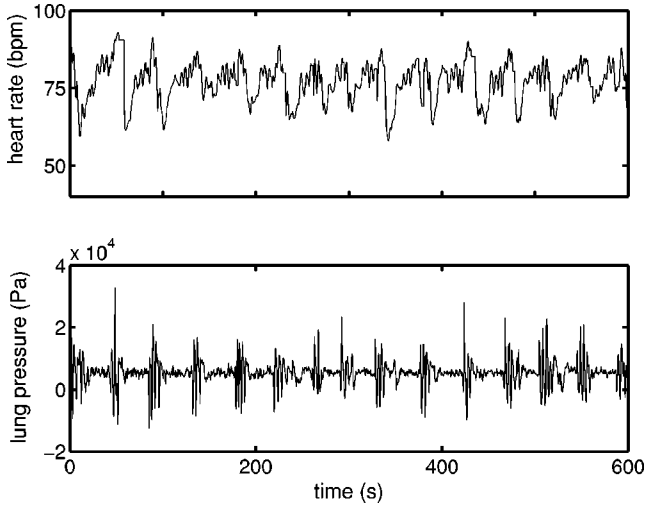

FIG. 3. Time series of the heart RR (upper) and breath signal (lower) of a patient suffering sleep apnea. Data sampled at 2 Hz.

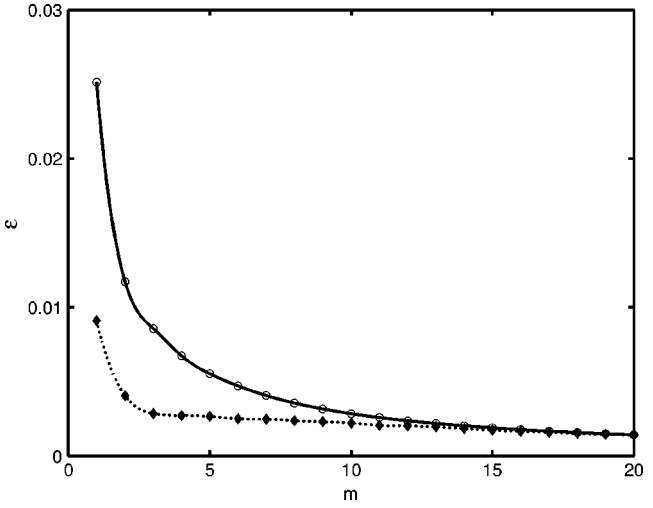

FIG. 4. ϵ_x (diamonds, lower curve) and ϵ_y (open circles, upper curve) are plotted vs m, for the sleep apnea example.

rather large value. It is worth stressing that the value $m=5$, at which peaks occur, is reasonable. Indeed in terms of frequency it corresponds to 0.4 Hz; it is well known that the high-frequency band (0.15–0.45 Hz) is characteristic of the respiratory rhythm. These data have been already analyzed in [8], measuring the rate of information flow (transfer entropy), and a stronger flow of information from the heart rate to the breath rate was found. In this example, the rate of information flow entropy and Granger nonlinear causality give consistent results: both these quantities, in the end, measure the departure from the generalized Markov property [8]

$$P(x|\mathbf{X}) = P(x|\mathbf{X},\mathbf{Y}),$$

$$P(y|\mathbf{Y}) = P(y|\mathbf{X},\mathbf{Y}). \tag{13}$$

As another physiological application we consider now rat EEG signals from right and left cortical intracranial electrodes, employed in the study of the pathophysiology of epi-

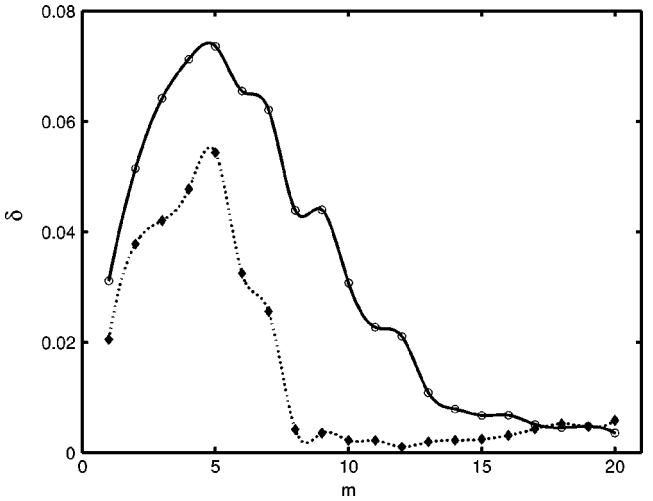

FIG. 5. δ_1 (diamonds, lower curve) and δ_2 (open circles, upper curve) are plotted vs m, for the sleep apnea example.

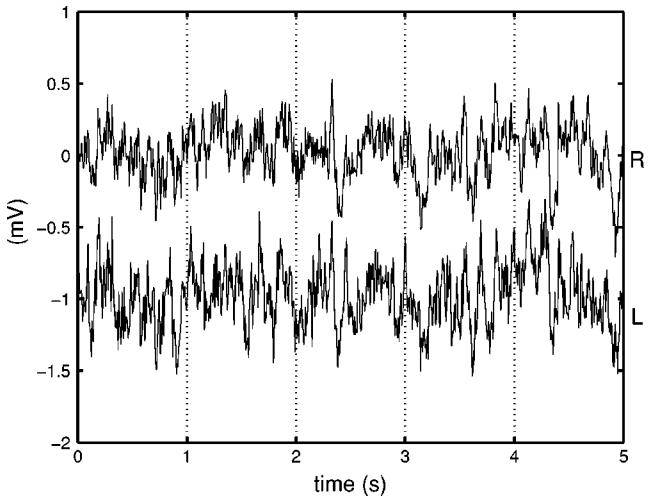

FIG. 6. Normal Rat EEG signals from right and left cortical intracranial electrodes. For a better visualization, left signals are plotted with an offset. Sampling rate is 200 Hz.

lepsy and already analyzed in [7]. In Fig. 6 the normal EEG signals (example A in [7]) from the rat is depicted. In Fig. 7 we depict the EEG signal from the same rat after unilateral lesion in the rostral pole of the reticular thalamic nucleus (example B in [7]): in this case spike discharges can be seen, due to local synchronization of neurons activity in the neighborhood of the electrode at which the signal was recorded. We remark that, as epilepsy is related to abnormal synchronization in the brain, spikes are usually considered as a landmark of epileptic activity. In order to analyze these recordings by Granger causality, we fix $\sigma=0.6$, $n=30$ and vary m in $\{1,2,\ldots,20\}$. In Fig. 8 we depict ϵ_x (x representing right EEG) and ϵ_y (y representing left EEG) versus m, while in Fig. 9 we depict the quantities δ_1 and δ_2, versus m. The pattern in Fig. 9 shows a slight asymmetry; i.e., the influence of the right channel on the left one seems to be slightly stronger than the reverse. The directionality index, evaluated

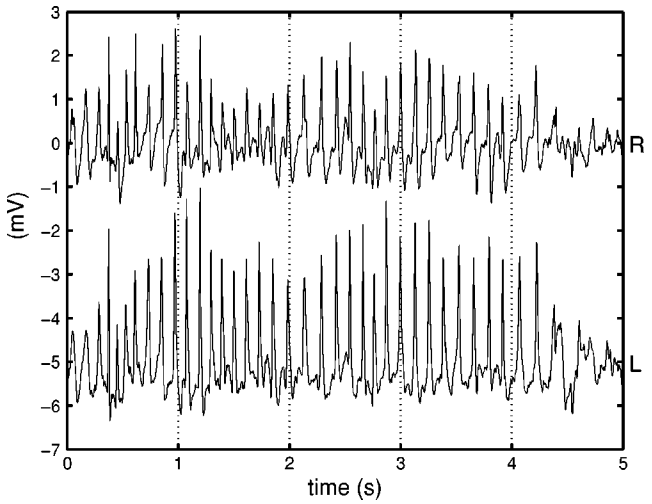

FIG. 7. Rat EEG signals from right and left cortical intracranial electrodes, after lesion. For a better visualization, left signals are plotted with an offset. Sampling rate is 200 Hz.

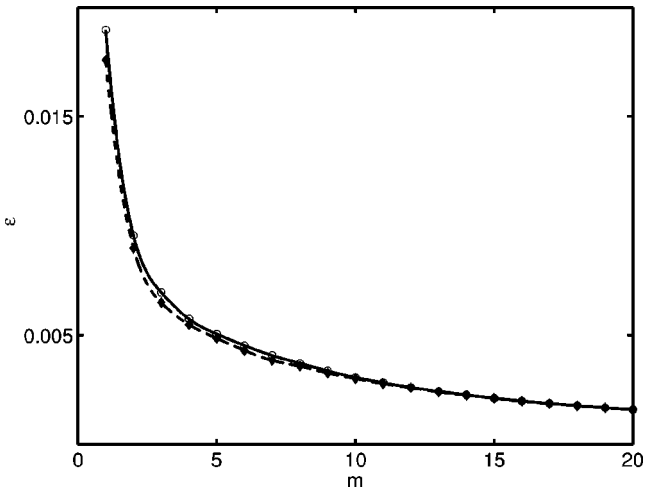

FIG. 8. ϵ_x (diamonds) and ϵ_y (open circles) are plotted vs m, for the EEG example A, with $n=30$. Curves are almost indistinguishable. The m value leading to the best model corresponds to the knee of the curves.

in correspondence of $m=5$, is $D=0.14$. Similar results are obtained varying n from 20 to 50. On the other hand, in the case of example B, the asymmetry is much more pronounced. In Fig. 10 we depict ϵ_x and ϵ_y, versus m, in this case for $n=30$. In Fig. 11, δ_1 and δ_2 versus m are depicted: in this figure the pattern is clearly suggesting that y is driving x. In other words, after the lesion the influence of the left signal on the right one is stronger and the peaks are now located at $m=4$. The directionality index, evaluated in correspondence of $m=4$, is now $D=-0.33$. Also in this case the results are found to be stable with respect to variations of n. Since example B is designed to mimic epileptic seizures, the pattern we find suggests that the focus is on the left side. Comparing with the analysis reported in [7], our analysis is in agreement with those from H and N measures of nonlinear interdependence (see [7]), which detected the same directions of asymmetry in examples A and B. Stronger interdependence in example B with respect to example A, like our method sug-

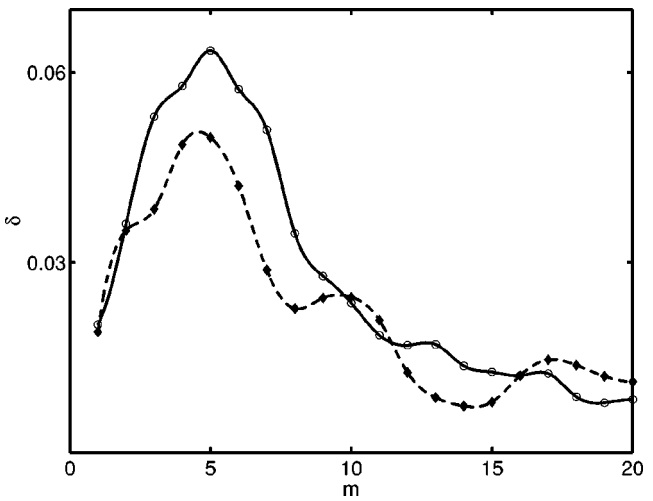

FIG. 9. δ_1 (diamonds) and δ_2 (open circles) are plotted vs m, for the EEG example A, with $n=30$.

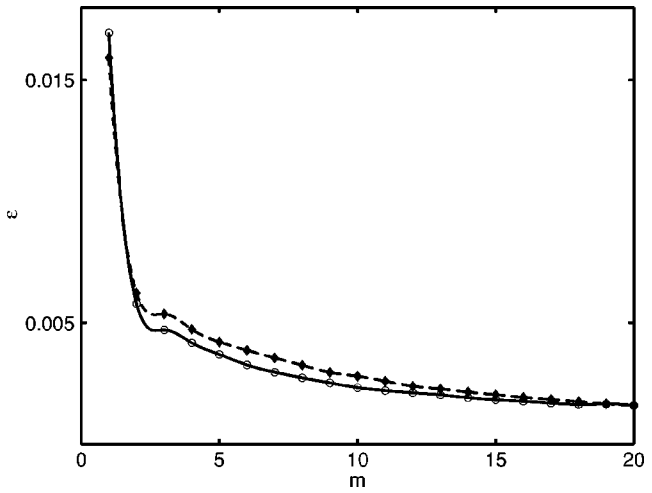

FIG. 10. ϵ_x (diamonds) and ϵ_y (open circles) are plotted vs m, for the EEG example B, with $n=30$.

gests, was also detected in [7]. We conclude this subsection stressing that our results show that the value of the directionality index D may in some cases be very sensitive to statistical fluctuations, especially when the interdependence is weak. Also other quantities, like $c_{1\text{-}2}$ or $\delta_{1\text{-}2}$, must then be taken into account to assess the Granger causality between two time series.

IV. CONCLUSIONS

The components of complex systems in nature rarely display a linear interdependence of their parts: identification of their causal relationships provides important insights into the underlying mechanisms. Among the variety of methods which have been proposed to handle this important task, a major approach was proposed by Granger [1]. It is based on improvement of the predictability of one time series due to knowledge of the second time series: it is appealing for its general applicability, but is restricted to linear models. While extending the Granger approach to the nonlinear case, on the one hand, one would like to have the most accurate modeling of the bivariate time series; on the other hand, the goal is to quantify how much knowledge of the other time series counts to reach this accuracy. Our analysis is rooted in the fact that any nonlinear modeling of data, suitable to study causality, should satisfy the property P1, described in Sec. II. It is clear that this property sets a limit on the accuracy of the model; we have proposed a class of nonlinear models which satisfy P1 and constructed an RBF-like approach to nonlinear Granger causality. Its performances, in a simulated case and real physiological applications, have been presented. We conclude remarking that use of this definition of nonlinear causality may lead to the discovery of genuine causal structures via data analysis, and validate the results that the analysis has to be accompanied by a substantive theory.

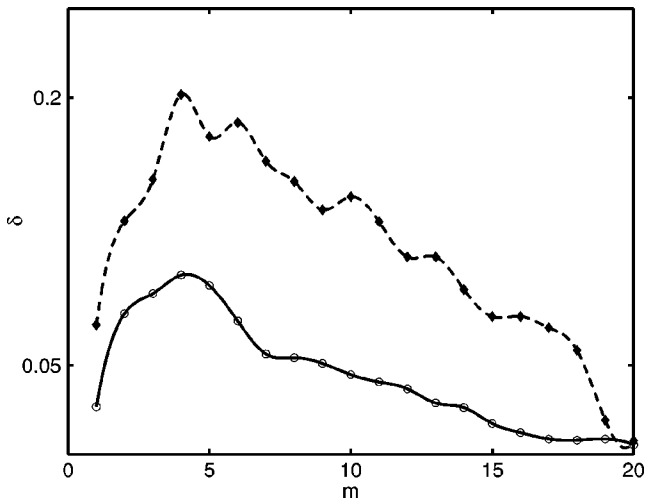

FIG. 11. δ_1 (diamonds) and δ_2 (open circles) are plotted vs m, for the EEG example B, with $n=30$. Note that the values of δ variables are larger, in this case, compared to values in Fig. 9.

ACKNOWLEDGMENTS

The authors thank Giuseppe Nardulli and Mario Pellicoro for useful discussions about causality.

APPENDIX: REGULARIZATION THEORY

We show how the choice of functions (8) arise in the frame of regularization theory. Let z be a function of $\mathbf{X}$ and $\mathbf{Y}$. We assume that z is the sum of a term depending solely on $\mathbf{X}$ and one depending on $\mathbf{Y}$: $z(\mathbf{X},\mathbf{Y})=f(\mathbf{X})+g(\mathbf{Y})$. We also assume knowledge of the values of f and g at points $\{\tilde{\mathbf{X}}^\rho,\tilde{\mathbf{Y}}^\rho\}_{\rho=1,\ldots,n}$:

$$f(\tilde{\mathbf{X}}^\rho)=f^\rho,\ \ \rho=1,\ldots,n,$$

$$g(\tilde{\mathbf{Y}}^\rho)=g^\rho,\ \ \rho=1,\ldots,n. \tag{A1}$$

Let us denote $\hat{K}(\vec{\omega})$ the Fourier transform of $K(\vec{r})=\exp(-r^2/2\sigma^2)$. The following functional is a measure of the smoothness of $z(\mathbf{X},\mathbf{Y})$:

$$\mathcal{S}[z]=\int d\vec{\omega}\frac{|\hat{f}(\vec{\omega})|^2+|\hat{g}(\vec{\omega})|^2}{\hat{K}(\vec{\omega})}. \tag{A2}$$

Indeed it penalizes functions with relevant contributions from high-frequency modes. Variational calculus shows that the function that minimize $\mathcal{S}$ under the constraints (A1) is given by

$$z=\sum_{\rho=1}^{n}\mu_\rho K(\mathbf{X}-\tilde{\mathbf{X}}^\rho)+\sum_{\rho=1}^{n}\lambda_\rho K(\mathbf{Y}-\tilde{\mathbf{Y}}^\rho), \tag{A3}$$

where $\{\mu\}$ and $\{\lambda\}$ are tunable Lagrange multipliers to solve Eqs. (A1). Hence the model (4)–(8) corresponds to the class of the smoothest functions, the sum of a term depending on **X** and a term depending on **Y**, with assigned values on a set of n points.

The standard RBF modeling of the bivariate time series [11], to be compared with model (4), is the following:

$$x = \sum_{\rho=1}^{n'} w_1^\rho K(\mathbf{Z} - \tilde{\mathbf{Z}}^\rho),$$

$$y = \sum_{\rho=1}^{n'} w_2^\rho K(\mathbf{Z} - \tilde{\mathbf{Z}}^\rho), \tag{A4}$$

where $\mathbf{Z} = (\mathbf{X}\ \mathbf{Y})$ is the vector obtained appending **X** and **Y**, $\{\tilde{\mathbf{Z}}^\rho\}$ are obtained by clustering the **Z** data, and $\mathbf{w}_{1-2}$ are determined by the least-squares method. In general, model (A4) does not satisfy property P1; hence, it is not suited to evaluate causality.

[1] C. W. J. Granger, Econometrica **37**, 424 (1969).

[2] J. J. Ting, Physica A **324**, 285 (2003).

[3] P. Tass *et al.*, Phys. Rev. Lett. **81**, 3291 (1998); M. Le van Quyen *et al.*, Brain Res. **792**, 24 (1998); E. Rodriguez *et al.*, Nature (London) **397**, 430 (1999).

[4] C. Ludwig, Arch. Anat. Physiol. **13**, 242 (1847); C. Schafer *et al.*, Phys. Rev. E **60**, 857 (1999); M. G. Rosenblum *et al.*, *ibid.* **65**, 041909 (2002).

[5] S. Akselrod *et al.*, Am. J. Physiol. Heart Circ. Physiol. **249**, H867 (1985); G. Nollo *et al.*, *ibid.* **283**, H1200 (2002).

[6] S. J. Schiff *et al.*, Phys. Rev. E **54**, 6708 (1996); J. Arnhold *et al.*, Physica D **134**, 419 (1999); R. Quian Quiroga *et al.*, Phys. Rev. E **61**, 5142 (2000); M. Wiesenfeldt, U. Parlitz, and W. Lauterborn, Int. J. Bifurcation Chaos Appl. Sci. Eng. **11**, 2217 (2001).

[7] R. Quian Quiroga *et al.*, Phys. Rev. E **65**, 041903 (2002).

[8] T. Schreiber, Phys. Rev. Lett. **85**, 461 (2000).

[9] M. Palus *et al.*, Phys. Rev. E **63**, 046211 (2001).

[10] F. R. Drepper, Phys. Rev. E **62**, 6376 (2000); M. G. Rosenblum *et al.*, *ibid.* **64**, 045202(R) (2001); M. Palus and A. Stefanovska, *ibid.* **67**, 055201(R) (2003).

[11] C. M. Bishop, *Neural Networks for Pattern Recognition* (Oxford University Press, New York, 1995).

[12] Usually both times series are normalized in the preprocessing stage; i.e., they are linearly transformed to have zero mean and unit variance.

[13] Y. Chen *et al.*, Phys. Lett. A **324**, 26 (2004).

[14] J. D. Farmer and J. J. Sidorowich, Phys. Rev. Lett. **59**, 845 (1987).

[15] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms* (Plenum, New York, 1981).

[16] T. Poggio and F. Girosi, Science **247**, 978 (1990).

[17] V. Vapnik, *Statistical Learning Theory* (Wiley, New York, 1998).

[18] Machine means *algorithm which learns from data* in the machine learning community.

[19] In the general case, leave-one-out (Loo) error $E_{loo}[f]$ provides a better estimate of the generalization error of f (Luntz-Brailovsky theorem) than the empirical risk, given a finite number of training data. Loo error is defined as the error variance when the prediction for the kth pattern is made using the model trained on the $M-1$ other patterns; it needs M predictors to be trained, where M is the cardinality of the data set. Hence, a Loo-error estimation is unfeasible to compute for large training sets. For linear predictors, like the ones we consider in this paper, the empirical risk is already a good estimate, due to the low complexity of these machines.

[20] H. F. von Bremen *et al.*, Physica D **101**, 1 (1997).

[21] http://www. physionet.org/

[22] F. Roux *et al.*, Am. J. Med. **108**, 396 (2000).

[23] H. W. Duchna *et al.*, Somnologie **7**, 101 (2003).

[24] These results may also be due to coupling of the two signals to a common external driver.