Kernel Method for Nonlinear Granger Causality

Daniele Marinazzo, Mario Pellicoro, and Sebastiano Stramaglia

Dipartimento Interateneo di Fisica, Università di Bari, Italy
Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Italy
TIRES-Center of Innovative Technologies for Signal Detection and Processing, Università di Bari, Italy

Important information on the structure of complex systems can be obtained by measuring to what extent the individual components exchange information among each other. The linear Granger approach, to detect cause-effect relationships between time series, has emerged in recent years as a leading statistical technique to accomplish this task. Here we generalize Granger causality to the nonlinear case using the theory of reproducing kernel Hilbert spaces. Our method performs linear Granger causality in the feature space of suitable kernel functions, assuming arbitrary degree of nonlinearity. We develop a new strategy to cope with the problem of overfitting, based on the geometry of reproducing kernel Hilbert spaces. Applications to coupled chaotic maps and physiological data sets are presented.

Experiments in many fields of science provide a time series of simultaneously recorded variables. The analysis of the synchronization between time series [1] is an important tool to study communications between different components of complex systems. In many systems, however, it is important not only to detect synchronized states, but also to identify cause-effect (drive-response) relationships between components [2,3]. Information-theoretic approaches to causality are based on the estimation of entropy and mutual information [4–6], a hard numerical problem when conditioning with respect to a large number of variables is to be done. Another major approach to analyzing causality between two time series has been proposed by Granger [7]: if the prediction error of the first time series is reduced by including measurements from the second one in the linear regression model, then the second time series is said to have a causal influence on the first one. This linear frame for measuring causality has been widely applied in many fields, including rheochaos [8] neurophysiology [9], economy [10], and climatology [11]. The importance of Granger causality is the suggestion to use prediction (and tools from learning theory [12] in particular) to measure the amount of information exchanged by two (sub)systems; it is worth mentioning that also a measure of self-organization, rooted in optimal predictors, has been recently proposed [13]. Some attempts to extend Granger causality to the nonlinear case have been recently proposed [14]. The main problem of all approaches is the detection of false causalities [15], which may arise due to overfitting of the learning scheme.

The purpose of this work is to present a novel approach that measures Granger causality of time series, assuming an arbitrary degree of nonlinearity, while controlling overfitting, and thus avoiding the problem of false causalities. To this aim we exploit the properties of kernel machines, the state of the art in learning models [16].

We start describing the connection between Granger causality and information-theoretic approaches like the transfer entropy T_E in [4]. Let $\{\xi_n\}_{n=1,\dots,N+m}$ be a time series that may be approximated by a stationary Markov process of order m, i.e., $p(\xi_n|\xi_{n-1},\dots,\xi_{n-m}) = p(\xi_n|\xi_{n-1},\dots,\xi_{n-m-1})$. We will use the shorthand notation $X_i = (\xi_i,\dots,\xi_{i+m-1})^\top$ and $x_i = \xi_{i+m}$, for $i = 1,\dots,N$, and treat these quantities as N realizations of the stochastic variables X and x. The minimizer of the risk functional, $R[f] = \int dX dx (x - f(X))^2 p(X, x)$, represents the best estimate of x, given X, and corresponds [17] to the regression function $f^*(X) = \int dx p(x|X)x$. Now, let $\{\eta_n\}_{n=1,\dots,N+m}$ be another time series of simultaneously acquired quantities, and denote $Y_i = (\eta_i,\dots,\eta_{i+m-1})^\top$. The best estimate of x, given X and Y, is now $g^*(X, Y) = \int dx p(x|X, Y)x$. If the generalized Markov property holds, i.e.,

$$p(x|X, Y) = p(x|X), \tag{1}$$

then $f^*(X) = g^*(X, Y)$ and the knowledge of Y does not improve the prediction of x. T_E [4] is a measure of the violation of (1); it follows that Granger causality implies a nonzero transfer entropy.

Because of the finiteness of N, the risk functional cannot be evaluated; we consider the empirical risk $ER[f] = \sum_{i=1}^{N}(x_i - f(X_i))^2$, and the search for the minimum of ER is constrained in a suitable functional space, called hypothesis space; the simplest choice is the space of all linear functions, corresponding to linear regression. In the following we propose a geometrical description of linear Granger causality. For each $\alpha \in \{1,\dots,m\}$, the samples of the αth component of X form a vector $\mathbf{u}_\alpha \in \Re^N$; without loss of generality we assume that each $\mathbf{u}_\alpha$ has zero mean and that $\mathbf{x} = (x_1,\dots,x_N)^\top$ is normalized and zero mean. We denote $\tilde{x}_i$ the value of the linear regression of x versus X, evaluated at X_i. The vector $\tilde{\mathbf{x}} = (\tilde{x}_1,\dots,\tilde{x}_N)^\top$ can be

obtained as follows. Let $H \subseteq \Re^N$ be the span of $\mathbf{u}_1, \ldots, \mathbf{u}_m$; then $\tilde{\mathbf{x}}$ is the projection of $\mathbf{x}$ on H. In other words, calling P the projector on the space H, we have $\tilde{\mathbf{x}} = P\mathbf{x}$. Moreover, the prediction error, given X, is $\epsilon_x = ||\mathbf{x} - \tilde{\mathbf{x}}||^2 = \mathbf{1} - \tilde{\mathbf{x}}^\top \tilde{\mathbf{x}}$. Calling $\mathbf{X}$ the $m \times N$ matrix having vectors $\mathbf{u}_\alpha$ as rows, H coincides with the range of the $N \times N$ matrix $\mathbf{K} = \mathbf{X}^\top \mathbf{X}$.

Using both X and Y, the values of the linear regression form the vector $\tilde{\mathbf{x}}' = P'\mathbf{x}$, P' being the projector on the space $H' \subseteq \Re^N$, spanned by the $\mathbf{u}_1, \ldots, \mathbf{u}_m$ and the components of Y $\mathbf{v}_1, \ldots, \mathbf{v}_m$ (assumed to be zero mean). H' is the range of the matrix $\mathbf{K}' = \mathbf{Z}^\top \mathbf{Z}$, where $\mathbf{Z}$ is the $2m \times N$ matrix with vectors $\mathbf{u}_\alpha$ and $\mathbf{v}_\alpha$ as rows. The prediction error is now $\epsilon_{xy} = ||\mathbf{x} - \tilde{\mathbf{x}}'||^2 = 1 - \tilde{\mathbf{x}}'^\top \tilde{\mathbf{x}}'$. We now note that $H \subseteq H'$; hence, $H' = H \oplus H^\perp$. The last formula shows geometrically the enlargement of the hypothesis space, due to the inclusion of the Y variables. Calling $P^\perp$ the projector on $H^\perp$, we have $\epsilon_{xy} = \epsilon_x - ||P^\perp \mathbf{x}||^2$, and the linear Granger causality index reads

$$\delta(Y \to X) = \frac{\epsilon_x - \epsilon_{xy}}{\epsilon_x} = \frac{||P^\perp \mathbf{x}||^2}{1 - \tilde{\mathbf{x}}^\top \tilde{\mathbf{x}}}. \qquad (2)$$

Linear Granger causality is usually assessed according to well-known test statistics; see, e.g., [2]. Instead of assessing the presence (or not) of causality by means of a single statistical test, and in view of the nonlinear extension, we introduce a causality index which by construction is not affected by overfitting. We observe that $H^\perp$ is the range of the matrix $\tilde{\mathbf{K}} = \mathbf{K}' - P\mathbf{K}' - \mathbf{K}'P + P\mathbf{K}'P$. Hence the natural choice of the orthonormal basis in $H^\perp$ is the set of the eigenvectors, with nonvanishing eigenvalue, of $\tilde{\mathbf{K}}$. Calling $\mathbf{t}_1, \ldots, \mathbf{t}_m$ these eigenvectors, we have $||P^\perp \mathbf{x}||^2 = \sum_{i=1}^m r_i^2$, where r_i is the Pearson's correlation coefficient of $\mathbf{x}$ and $\mathbf{t}_i$. To avoid false causalities, we first evaluate, by Student's t test, the probability π_i that r_i is due to chance, assuming $\mathbf{x}$ and $\mathbf{t}_i$ normal. Since we are dealing with multiple comparison, we use the Bonferroni correction to select the eigenvectors $\mathbf{t}_{i'}$, correlated with $\mathbf{x}$, with expected fraction of false positive equal to 0.05. Then we calculate a new causality index by summing only over the $\{r_{i'}\}$ which pass the Bonferroni test, thus obtaining what we call filtered linear Granger causality index:

$$\delta_F(Y \to X) = \frac{\sum_{i'} r_{i'}^2}{1 - \tilde{\mathbf{x}}^\top \tilde{\mathbf{x}}}. \qquad (3)$$

Exchanging the roles of the two time series, we may evaluate the causality index in the opposite direction $\delta_F(X \to Y)$.

The formulation of linear Granger causality, above described, allows an efficient generalization to the nonlinear case using methods of the theory of reproducing kernel Hilbert spaces [16]. Let us first deal with the problem of predicting x using the knowledge of X. Given a kernel function k, with spectral representation $k(X, X') = \sum_a \lambda_a \Psi_a(X) \Psi_a(X')$, we consider H, the range of the $N \times N$ Gram matrix $\mathbf{K}$ with elements $K_{ij} = k(X_i, X_j)$. As in the

linear case, we calculate $\tilde{\mathbf{x}}$, the projection of $\mathbf{x}$ onto H. Because of the spectral representation of k, $\tilde{\mathbf{x}}$ coincides with the linear regression of $\mathbf{x}$ in the feature space spanned by $\sqrt{\lambda_a} \Psi_a$, the eigenfunctions of k; the regression is nonlinear in the original variables. We remark that H corresponds to the functional space where well-known methods, like support vector machines and kernel ridge regression, search for the regressor [16].

While using both X and Y to predict x, we append X and Y variables to construct the Z variable with samples $Z_i = (X_i Y_i)^\top$; then we evaluate the Gram matrix $\mathbf{K}'$ with elements $K'_{ij} = k(Z_i, Z_j)$. The regression values now form vector $\tilde{\mathbf{x}}'$ equal to the projection of $\mathbf{x}$ on H', the range of $\mathbf{K}'$. In this work we consider two choices of the kernel (see the discussion in [18]): the inhomogeneous polynomial (IP) of integer order p: $k_p(X, X') = (1 + X^\top X')^p$, and the Gaussian: $k_\sigma(X, X') = \exp(-\frac{(X - X')^\top (X - X')}{2\sigma^2})$, whose complexity depends on the scale parameter σ.

First we consider the IP kernel. In this case the eigenfunctions Ψ_a are all the monomials, in the input variables, up to the pth degree. In this case $H \subseteq H'$, and we can proceed as in the linear case, decomposing $H' = H \oplus H^\perp$ and calculating $\tilde{\mathbf{K}} = \mathbf{K}' - P\mathbf{K}' - \mathbf{K}'P + P\mathbf{K}'P$. Along the same lines as those described in the linear case, we may construct the filtered Granger causality taking into account only the eigenvectors of $\tilde{\mathbf{K}}$ which pass the Bonferroni test. We discuss some examples of application of our method with IP kernel. First we consider two unidirectionally coupled noisy logistic maps:

$$\begin{aligned} x_{n+1} &= 1 - ax_n^2 + s\tau_n, \\ y_{n+1} &= (1 - e)(1 - ay_n^2) + e(1 - ax_n^2) + s\gamma_n; \end{aligned} \qquad (4)$$

$\{\tau\}$ and $\{\gamma\}$ are unit variance Gaussianly distributed noise terms (the parameter s determines their relevance), $a = 1.8$ and $e \in [0, 1]$ represents the coupling $x \to y$. In the noise-free case ($s = 0$), a transition to complete synchronization [1] occurs at $e = 0.37$. Varying e and s, we have considered runs of N iterations, after a transient of 10^3, and evaluated δ_F, using the kernel with $p = 2$, in both directions. We find that $\delta_F(Y \to X)$ is zero for all values of e, s, and N. On the other hand $\delta_F(X \to Y)$ is zero at e smaller than a threshold e_c; see Fig. 1. $\delta_F(X \to Y)$ is zero also at complete synchronization, as there is no information transfer in this regime. As noted in [15], the causal relation can be inferred only when the coupling is not large enough to let full synchronization emerge, or when the synchronized state is frequently perturbed by internal or external noise driving the system out of the synchronized state. Indeed, at fixed $e > 0.37$, $\delta_F(X \to Y)$ is zero until s reaches a threshold s_c. Both e_c and s_c scale as $N^{-0.5}$, as expected, see Fig. 2.

As another simulated example, we consider two unidirectionally coupled Henon maps:
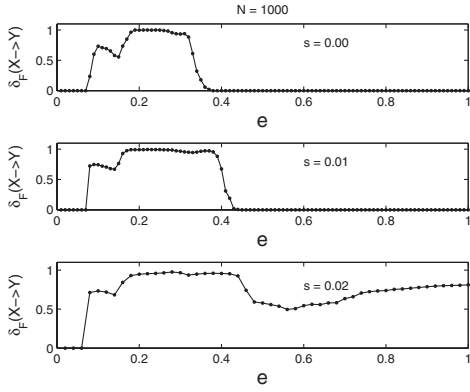
FIG. 1.   The filtered causality index, for the coupled maps, is displayed versus e for three values of s. The inhomogeneous polynomial kernel with $p = 2$ is used, and $m = 1$.

$$x_n = 1.4 - x_{n-1}^2 + 0.3\hat{x}_{n-1}; \qquad \hat{x}_n = x_{n-1};$$

$$y_n = 1.4 + 0.3\hat{y}_{n-1} - 0.7y_{n-1}^2 - 0.3x_{n-1}y_{n-1};$$

$$\hat{y}_n = y_{n-1};$$

and analyze the causality between time series $\{x\}$ and $\{y\}$; by construction, x is driving y. Using $m = 2$ and IP kernel with various values of p, on runs of length $N = 1000$, we correctly find that the causality $y \rightarrow x$ is always zero while the causality $x \rightarrow y$ is nonzero and maximal at $p = 2$, the interaction being quadratic (Table I).

A real example consists in rat electroencephalogram (EEG) signals from right (R) and left (L) cortical intra-cranial electrodes, employed in the study of the patho-physiology of epilepsy and already analyzed in [19]. We analyze both the normal EEG signals and the EEG signals from the same rat after unilateral lesion in the rostral pole of the reticular thalamic nucleus, when spike discharges are observed due to local synchronization of neurons activity in the neighborhood of the electrode at which the signal was recorded. In Fig. 3 top, the indices δ_F, for the normal EEG signals of the rat, are depicted for $p = 1, 2, 3, 4$. We find zero causality in the direction $L \rightarrow R$ for all p, and a small causality $R \rightarrow L$ only at $p = 1$. After the unilateral lesion (Fig. 3 bottom) we find that causality $R \rightarrow$
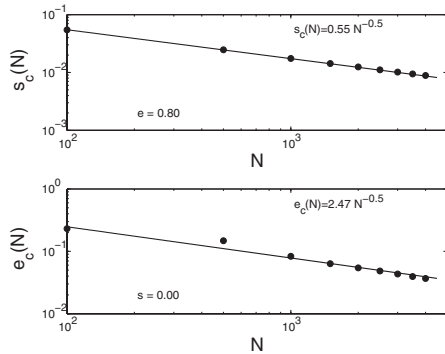

FIG. 2.   Scaling of the critical values of e and s (see the text) with N. e_c also has a weak dependence on s.

TABLE I.   Causalities $x \rightarrow y$ for coupled Henon's maps.

p	1	2	3	4	5
δ_F	0.04	0.88	0.81	0.66	0.41

L is almost unchanged, while a relevant $L \rightarrow R$ causality now appears at $p = 1$ and (smaller) at $p = 2$. A more conservative statistical procedure, in situations where the value of p is not known *a priori*, is to apply, at each p, the Bonferroni's correction corresponding to the total number of comparisons, in this case 91 ($= 2 + 9 + 25 + 55$); using this correction, the causality $R \rightarrow L$ becomes zero in both cases and for all p, while the causality $L \rightarrow R$ remains unchanged and equal to the values depicted in Fig. 3. The results reported in [19] are qualitatively consistent with our findings, indeed the same directions of asymmetry are found in the two analyses, but our approach allows us to make more sharp and precise statements about the causality relationships between the two EEG signals: the only statistically robust causality relationship is $L \rightarrow R$ after the lesion. Moreover, as the maximum of $\delta_F(L \rightarrow R)$ occurs at $p = 1$, our analysis seems to suggest that in this experiment the information transfer mechanism is essentially linear [20].

Turning to consider the Gaussian kernel, the condition $H \subseteq H'$ does not necessarily hold and some differences in the approach are in order. In this case we call H the span of the eigenvectors of $\mathbf{K}$ whose eigenvalue is not smaller than $\mu\lambda_{\max}$, where $\lambda_{\max}$ is the largest eigenvalue of $\mathbf{K}$ and μ is a small number (we use $\mu = 10^{-6}$). We calculate $\tilde{\mathbf{x}} = P\mathbf{x}$, where P is the projector on H. After evaluating the Gram matrix $\mathbf{K}'$, the following matrix is considered:

$$\mathbf{K}^* = \sum_{i=1}^{m_2} \rho_i \mathbf{w}_i \mathbf{w}_i^\top, \qquad (5)$$

where $\{\mathbf{w}\}$ are the eigenvectors of $\mathbf{K}'$, and the sum is over the eigenvalues $\{\rho_i\}$ not smaller than μ times the largest eigenvalue of $\mathbf{K}'$. Then we evaluate $\tilde{\mathbf{K}} = \mathbf{K}^* - \mathbf{P}\mathbf{K}^* - \mathbf{K}^*\mathbf{P} + \mathbf{P}\mathbf{K}^*\mathbf{P}$, and denote $P^\perp$ the projector onto the range
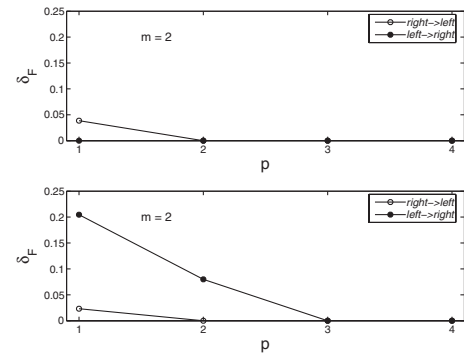

FIG. 3.   The filtered causality indexes, for the rat EEG signals before (top) and after the lesion (bottom), is displayed versus p, the order of the inhomogeneous polynomial kernel.
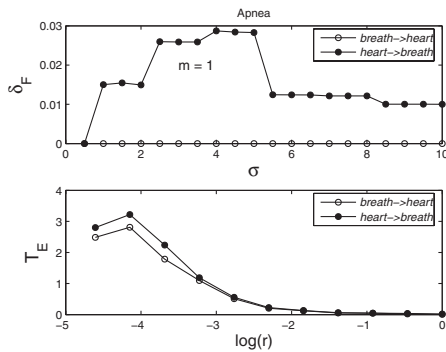
FIG. 4. (Top) The filtered causality indexes, for the physionet data set, are displayed versus σ, the width of the Gaussian kernel. (Bottom) Transfer entropies versus r, the length scale.

of $\tilde{K}$. The filtered Granger causality index for Gaussian kernels is then constructed as in the previous cases. As another real example, we consider time series of heart rate (H) and breath rate (B) of a sleeping human suffering from sleep apnea (ten minutes from data set B of the Santa Fe Institute time series contest held in 1991, available in the Physionet data bank [21]). Using IP kernels, we find unidirectional causality $H \to B$; its strength increases with the order p of the kernel, from $\delta_F = 0.01$ at $p = 1$ to $\delta_F = 0.03$ at $p = 5$. These findings confirm the strongly nonlinear nature of the interaction between heart and respiration signals in sleep apnea syndrome [22], which is evident also using the Gaussian kernel and varying σ, as depicted in Fig. 4 top. No causality $B \to H$ is found to be significative, while nonzero causality $H \to B$ is found for $\sigma \gtrsim 1$. Note that the causality index vanishes, by construction, at small σ and at large σ, because in both limits the kernel matrix tends to be constant (0 and 1, respectively). In Fig. 4 bottom the bivariate time series is analyzed by means of the transfer entropy [4]. It is interesting to compare the two approaches in this application. T_E is nonzero in both directions and shows a slightly stronger flow of information $H \to B$. Our approach recognizes, as significative, only the causality $H \to B$, thus revealing unidirectional drive-response relationship in the sleep apnea pathology.

In conclusion, we considered the problem of nonlinear coherence of signals, in particular, the detection of drive-response relationships. Exploiting the geometry of reproducing kernel Hilbert spaces we have introduced a filtered index which is able to measure cause-effect relationships with an arbitrary amount of nonlinearity, and is not affected by overfitting. The choice of the optimal value of m can be done using the standard cross validation scheme [16] or the embedding dimension [23]. Our method is equivalent to performing a linear Granger causality in the feature space of the kernel; hence, also in the nonlinear case, our approach continues to fulfill the good properties of linear models. The framework of Granger causality assumes stationarity of signals: further work should deal with the effects of nonstationarities on nonlinear estimates of cau-

salities (see [24] for a promising strategy in the linear case).

We expect that the proposed method will provide a statistically robust basis to assess nonlinear drive-response relationships in many fields of science, wherever collected data form time series; it works for deterministic and stochastic systems, provided that noise is not so high as to obscure the deterministic effects.

[1] A. Pikovsky, M. Rosenblum, and J. Kurths, *Synchronization, a Universal Concept in Nonlinear Sciences* (Cambridge University Press, Cambridge, U.K., 2001).

[2] K. Hlavackova-Schindler, M. Palus, M. Vejmelka, and J. Bhattacharya, Phys. Rep. **441**, 1 (2007).

[3] M. Lungarella, K. Ishiguro, Y. Kuniyoshi, and N. Otsu, Int. J. Bifurcation Chaos Appl. Sci. Eng. **17**, 903 (2007).

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

[5] M. Wiesenfeldt, U. Parlitz, and W. Lauterborn, Int. J. Bifurcation Chaos Appl. Sci. Eng. **11**, 2217 (2001).

[6] S. Frenzel and B. Pompe, Phys. Rev. Lett. **99**, 204101 (2007).

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

[8] R. Ganapathy *et al.*, Phys. Rev. E **75**, 016211 (2007).

[9] M. Kaminski, M. Z. Ding, W. A. Truccolo, and S. L. Bressler, Biol. Cybern. **85**, 145 (2001); K. J. Blinowska, R. Kus, and M. Kaminski, Phys. Rev. E **70**, 050902(R) (2004).

[10] G. Rodriguez, N. Rowe, J. Int. Money Finance **26**, 1174 (2007).

[11] U. Triacca, Theor. Appl. Climatol. **81**, 133 (2005).

[12] V. Vapnik, *Statistical Learning Theory* (John Wiley & Sons, Inc., New York, 1998).

[13] C. R. Shalizi, K. L. Shalizi, and R. Haslinger, Phys. Rev. Lett. **93**, 118701 (2004).

[14] Y. Chen *et al.*, Phys. Lett. A **324**, 26 (2004); P. F. Verdes, Phys. Rev. E **72**, 026222 (2005); D. Marinazzo *et al.*, Phys. Rev. E **73**, 066216 (2006).

[15] M. Palus and M. Vejmelka, Phys. Rev. E **75**, 056211 (2007); H. Nalatore and M. Z. Ding, Phys. Rev. E **75**, 031123 (2007).

[16] J. Shawe-Taylor and N. Cristianini, *Kernel Methods For Pattern Analysis* (Cambridge University Press, London, 2004).

[17] A. Papoulis, *Probability, Random Variables, and Stochastic Processes* (McGraw-Hill, New York, 1985).

[18] N. Ancona and S. Stramaglia, Neural Comput. **18**, 749 (2006).

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

[20] Another possible explanation is that, due to a high level of noise, nonlinear effects cannot be put in evidence in this experiment.

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

[22] J. A. Jo *et al.*, Ann. Biomed. Eng. **35**, 1425 (2007).

[23] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, U.K., 1997).

[24] M. Dhamala, G. Rangarajan, and M. Ding, Phys. Rev. Lett. **100**, 018701 (2008).