A GAUSSIAN PROCESS REGRESSION APPROACH FOR TESTING GRANGER CAUSALITY BETWEEN TIME SERIES DATA

P. O. Amblard$^{(1,2)}$, *O. J. J. Michel*$^{(2)}$

$^{(1)}$ Dept. of Math&Stat
The University of Melbourne, Australia
$^{(2)}$ GIPSA-lab, UMR CNRS 5216
Grenoble, France

C. Richard$^{(3)}$, *P. Honeine*$^{(4)}$

$^{(3)}$ Université de Nice Sophia-Antipolis, France
Institut Universitaire de France
$^{(4)}$ STMR, UMR CNRS 6279
Université de Technologie de Troyes, France

ABSTRACT

Granger causality considers the question of whether two time series exert causal influences on each other. Causality testing usually relies on prediction, *i.e.*, if the prediction error of the first time series is reduced by taking measurements from the second one into account, then the latter is said to have a causal influence on the former. In this paper, a nonparametric framework based on functional estimation is proposed. Nonlinear prediction is performed via the Bayesian paradigm, using Gaussian processes. Some experiments illustrate the efficiency of the approach.

Index Terms— Granger causality, functional estimation, Gaussian process, reproducing kernel

1. INTRODUCTION

Granger causality is an answer to the question of assessing possible influences between times series, and consists of determining whether and how two time series have influences on each other. This principle has been developed extensively in econometry. See, e.g., [1, 2, 3, 4] to cite some but a very few. It has also been used during the last decade within various fields as diverse as climatology [5] or neuroscience [6]. Surprisingly, it has only been rarely considered in the signal processing community [7, 8].

The notion of Granger causality relies on prediction and basically states that a signal x_t is a cause of a signal y_t if the prediction of y_t based on its past is improved when using the past of x_t also. A formal definition states that x_t does not cause y_t if and only if

$$p(y_t|y^{t-1}) = p(y_t|x^{t-1}, y^{t-1}) \tag{1}$$

where x^t denotes the whole past up to time instant t of the signal x_t. When other signals than x_t and y_t are observed, they must be taken into account in the definition, and the above Markov condition must be considered conditionally to these other observations. The conditional independence condition then writes $p(y_t|y^{t-1}, x^{t-1}, w^{t-1}) = p(y_t|y^{t-1}, w^{t-1})$. In the sequel we will consider that x_t and y_t are scalar signals, whereas w_t can be a multivariate signal. Testing for Granger causality was performed in several ways in the literature. For example, dependence measures were considered in [9, 10, 11] and others. Parametric modeling was used jointly with a statistical testing approach in [2, 3, 7, 8]. The linear-Gaussian case, which leads to detectors based on asymptotic variances of prediction errors, was usually considered within this framework. Some nonlinear models have also been considered [6].

In this paper, we propose a new nonparametric framework that relies on functional estimation and explicit prediction. As in [14], our approach exploits some concepts of Machine Learning associated with reproducing kernel Hilbert space theory. It however differs in the way how inference is handled. Here, forecasting is treated as a nonlinear regression problem using the Bayesian paradigm with Gaussian processes as priors [12]. Thus the solution can be then viewed as a time series prediction problem in the reproducing kernel Hilbert space defined by the covariance function used to model the prior. This has been recognized recently as an effective solution to nonlinear system identification problems, see *e.g.* [13] and references therein. This paper is organized as follows. The concept of Granger causality is briefly introduced in Section 2. Causality testing based on the framework of Gaussian processes is presented in Section 3. Section 4 reports some experiments. Finally, Section 5 sums up the results and describes the future works.

2. GRANGER CAUSALITY

Consider two scalar time series x_t and y_t, and the question of whether or not x_t (resp., y_t) causes y_t (resp., x_t). Let w_t be a third time series, possibly multidimensional, that models

P. O. Amblard is funded by a Marie Curie International Outgoing Fellowship from the European Community

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

some extra measurements and will be referred to as the side information. Two structural models are of interest:

Model 1:

$$\begin{aligned} x_t &= f_x^1(x^{t-1}, w^{t-1}) + \varepsilon_{x,t}^1 && (2)\\ y_t &= f_y^1(y^{t-1}, w^{t-1}) + \varepsilon_{y,t}^1 && (3)\end{aligned}$$

Model 2:

$$\begin{aligned} x_t &= f_x^2(x^{t-1}, y^{t-1}, w^{t-1}) + \varepsilon_{x,t}^2 && (4)\\ y_t &= f_y^2(x^{t-1}, y^{t-1}, w^{t-1}) + \varepsilon_{y,t}^2 && (5)\end{aligned}$$

The dynamical noises $\varepsilon_{x,t}^{(1,2)}$ and $\varepsilon_{y,t}^{(1,2)}$ are i.i.d. and, for a given model, not necessarily independent of each other. However, as developed in [2, 4], dependence between the dynamical noises is related to the notion of instantaneous causality. This topic is beyond the scope of this paper, and we further assume that the noises are independent from each other.

Clearly, the dynamical noises represent the innovation processes under each model. For example, $\varepsilon_{x,t}^1$ is the error obtained when optimally predicting x_t from its past and the past of w_t. Likewise, $\varepsilon_{x,t}^2$ is the prediction error of x_t from its past, and the past of y_t and w_t. Adopting Granger point of view on causality leads to the conclusion that the series y causes the series x if $\text{Var}[\varepsilon_{x,t}^2] < \text{Var}[\varepsilon_{x,t}^1]$, [1]. The aim is thus to derive the best predictor for both models and compare the power of the residuals. Parametric tests based on linear predictors have been extensively studied in the literature. Surprisingly, there are relatively few works considering possibly nonlinear relationships between time series. In this paper, we present a nonlinear framework based on Gaussian processes. See [12] for an overview. The specification of priors allows us to automatically reduce risks of overfitting, on the contrary of [14] that uses manual selection of the largest eigenvalues of some Gram matrices involved in the prediction processes.

3. TESTING GRANGER CAUSALITY WITH GAUSSIAN PROCESSES

Suppose we are given a training set[1]

$$\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i), (x_i, y_i)\}_{i=1,\ldots,N_\mathcal{D}} \tag{6}$$

where $\mathbf{x}_i$ and $\mathbf{y}_i$ denote the input vectors defined by

$$\begin{aligned} \mathbf{x}_i &= [x_{i-1},\, ...,\, x_{i-M_x+1}]^\top \\ \mathbf{y}_i &= [y_{i-1},\, \ldots,\, y_{i-M_y+1}]^\top \end{aligned} \tag{7}$$

and (x_i, y_i) the corresponding desired outputs. The column vector inputs $\mathbf{x}_i$ and $\mathbf{y}_i$, for all $N_\mathcal{D}$ cases, are aggregated in the design matrices X and Y. The targets x_i and y_i are collected in the vectors $\mathbf{f}_x$ and $\mathbf{f}_y$, and denoted by $f_{x,i}$ and $f_{y,i}$ for notational simplicity – See equation (9). We are interested in making inferences about the relationships (2)-(5), that is, the conditional distribution of the desired outputs given the input vectors. Suppose we are also given two matrices X^* and Y^* of input vectors of the following test set

$$\mathcal{T} = \{(\mathbf{x}_i^*, \mathbf{y}_i^*)\}_{i=1,\ldots,N_\mathcal{T}} \tag{8}$$

that have been aggregated.

3.1. Predicting with Gaussian processes

Functions f_x^1 and f_y^1 in Model 1 are assumed unknown. Both need to be identified in order to compute estimates $(f_{x,i}^*, f_{y,i}^*)$ at each test location $(\mathbf{x}_i^*, \mathbf{y}_i^*)$, that is,

$$\begin{aligned} f_{x,i}^* &\triangleq f_x^1(\mathbf{x}_i^*) \\ f_{y,i}^* &\triangleq f_y^1(\mathbf{y}_i^*) \end{aligned} \quad \text{and} \quad \begin{aligned} f_{x,i} &= f_x^1(\mathbf{x}_i) + \varepsilon_{x,i}^1 \\ f_{y,i} &= f_y^1(\mathbf{y}_i) + \varepsilon_{y,i}^1. \end{aligned} \tag{9}$$

The function values $f_{x,i}^*$ and $f_{y,i}^*$, for all $i = 1, \ldots, N_\mathcal{T}$, are collected in the vectors $\mathbf{f}_x^*$ and $\mathbf{f}_y^*$. Using the Gaussian process regression models introduced in [12], inference is performed on $(\mathbf{f}_x^*, \mathbf{f}_y^*)$ based on the Bayesian framework and priors on functions f_x^1 and f_y^1. The latter are considered as Gaussian random fields, and are completely characterized by their mean functions and their covariance functions. For the sake of simplicity, we will take the mean functions to be zero. The covariance functions of f_x^1 and f_y^1 will be denoted by $k_x^1(\mathbf{x}, \mathbf{x}')$ and $k_y^1(\mathbf{y}, \mathbf{y}')$, respectively. They depend on parameters that influence their shapes, and thus determine the sample path properties of the random fields f_x^1 and f_y^1.

In order to specify the distribution of $(f_{x,i}^*, f_{y,i}^*)$ conditionally to the input vectors of the training and test sets, we need to invoke the joint statistics of $f_x^1(\mathbf{x}_i)$, $f_y^1(\mathbf{y}_i)$, $f_x^1(\mathbf{x}_i^*)$ and $f_y^1(\mathbf{y}_i^*)$. We shall assume that the two random fields are independent, which implies that their covariance function $E[f_x^1(\mathbf{x}_i) f_y^1(\mathbf{y}_j)]$ is zero for all i, j. Let K_{xx}^1 be the $N_\mathcal{D} \times N_\mathcal{D}$ matrix whose (i,j)-th entry is $k_x^1(\mathbf{x}_i, \mathbf{x}_j)$. Let us denote by K_{yy}^1 the $N_\mathcal{D} \times N_\mathcal{D}$ matrix whose (i,j)-th entry is $k_y^1(\mathbf{y}_i, \mathbf{y}_j)$. Similarly, we introduce the following $N_\mathcal{D} \times N_\mathcal{T}$ matrices

$$\begin{aligned} [K_{xx^*}^1]_{ij} &= k_x^1(\mathbf{x}_i, \mathbf{x}_j^*) = [K_{x^*x}^1]_{ji} \\ [K_{yy^*}^1]_{ij} &= k_y^1(\mathbf{y}_i, \mathbf{y}_j^*) = [K_{y^*y}^1]_{ji} \end{aligned}$$

Finally, let $K_{x^*x^*}^1$ and $K_{y^*y^*}^1$ be the matrices with (i,j)-th entries $k_x^1(\mathbf{x}_i^*, \mathbf{x}_j^*)$ and $k_y^1(\mathbf{x}_i^*, \mathbf{x}_j^*)$, respectively.

The noises ε_x^1 and ε_y^1 are assumed to be Gaussian, i.i.d., and independent of each other. Let σ_x^2 and σ_y^2 be their variance. Then we may write the joint distribution of the target values and predicted values conditionally to the regressors as

$$P_1(\mathbf{f}_x, \mathbf{f}_y, \mathbf{f}_x^*, \mathbf{f}_y^* | X, Y, X^*, Y^*) = \mathcal{N}(0, \Sigma_1)$$

where Σ_1 is the covariance matrix given by

$$\Sigma_1 = \begin{pmatrix} K_{xx}^1 + \sigma_x^2 I & 0 & K_{xx^*}^1 & 0 \\ 0 & K_{yy}^1 + \sigma_y^2 I & 0 & K_{yy^*}^1 \\ K_{x^*x}^1 & 0 & K_{x^*x^*}^1 & 0 \\ 0 & K_{y^*y}^1 & 0 & K_{y^*y^*}^1 \end{pmatrix}$$

[1] Side information w_t in eq. (2) to (5) is omitted for the sake of clarity.

I is the identity matrix of appropriate dimension, and $\mathcal{N}(0, \Sigma)$ stands for the Gaussian distribution with mean vector 0 and covariance matrix Σ.

Predictive equations can then be derived. For Model 1, the posterior distribution $P_1(\mathbf{f}_x^*|X^*, X, \mathbf{f}_x)$ is given by

$$\begin{aligned} P_1(\mathbf{f}_x^*|X^*, X, \mathbf{f}_x) = \mathcal{N}\big(&K^1_{x^*x}(K^1_{xx} + \sigma_x^2 I)^{-1}\mathbf{f}_x, \\ &K^1_{x^*x^*} - K^1_{x^*x}(K^1_{xx} + \sigma_x^2 I)^{-1}K^1_{xx^*}\big) \end{aligned}$$

and the posterior distribution $P_1(\mathbf{f}_y^*|Y^*, Y, \mathbf{f}_y)$ is

$$\begin{aligned} P_1(\mathbf{f}_y^*|Y^*, Y, \mathbf{f}_y) = \mathcal{N}\big(&K^1_{y^*y}(K^1_{yy} + \sigma_y^2 I)^{-1}\mathbf{f}_y, \\ &K^1_{y^*y^*} - K^1_{y^*y}(K^1_{yy} + \sigma_y^2 I)^{-1}K^1_{yy^*}\big) \end{aligned}$$

Model 2 is a direct extension of the previous case with extended regression variables, as we consider now

$$\begin{array}{rcl} f^*_{x,i} & \triangleq & f^2_x(\mathbf{x}^*_i, \mathbf{y}^*_i) \\ f^*_{y,i} & \triangleq & f^2_y(\mathbf{x}^*_i, \mathbf{y}^*_i) \end{array} \quad \text{and} \quad \begin{array}{rcl} f_{x,i} & = & f^2_x(\mathbf{x}_i, \mathbf{y}_i) + \varepsilon^2_{x,i} \\ f_{y,i} & = & f^2_y(\mathbf{x}_i, \mathbf{y}_i) + \varepsilon^2_{y,i}. \end{array}$$

Let σ_x^2 and σ_y^2 be the variances of ε_x^2 and ε_y^2, which are assumed to be Gaussian, i.i.d., and independent of each other. The joint distribution of the target values and predicted values conditionally to the regressors has the following form

$$P_2(\mathbf{f}_x, \mathbf{f}_y, \mathbf{f}_x^*, \mathbf{f}_y^*|X, Y, X^*, Y^*) = \mathcal{N}(0, \Sigma_2)$$

with Σ_2 the covariance matrix defined by

$$\Sigma_2 = \begin{pmatrix} K^2_{zz} + \sigma_x^2 I & 0 & K^2_{zz^*} & 0 \\ 0 & K^2_{zz} + \sigma_y^2 I & 0 & K^2_{zz^*} \\ K^2_{z^*z} & 0 & K^2_{z^*z^*} & 0 \\ 0 & K^2_{z^*z} & 0 & K^2_{z^*z^*} \end{pmatrix}$$

where

$$\begin{aligned} &[K^2_{zz}]_{ij} = k^2_{xy}((\mathbf{x}_i, \mathbf{y}_i), (\mathbf{x}_j, \mathbf{y}_j)) \\ &[K^2_{zz^*}]_{ij} = k^2_{xy}((\mathbf{x}_i, \mathbf{y}_i), (\mathbf{x}^*_j, \mathbf{y}^*_j)) = [K^2_{z^*z}]_{ji} \end{aligned}$$

Posterior distributions of interest $P_2(\mathbf{f}_x^*|X^*, X, Y^*, Y, \mathbf{f}_x)$ and $P_2(\mathbf{f}_y^*|X^*, X, Y^*, Y, \mathbf{f}_y)$ are Gaussian with mean and covariance functions defined as for Model 1, substituting Gram matrices $K^1_{x^{(*)}x^{(*)}}$ and $K^1_{y^{(*)}y^{(*)}}$ by $K^2_{z^{(*)}z^{(*)}}$.

3.2. Evidences and measures of causality

A multitude of possible covariance functions, also called kernels, exists. A classic example of kernels is the radially Gaussian kernel $k(\mathbf{z}, \mathbf{z}') = \exp\big(-\|\mathbf{z} - \mathbf{z}'\|^2/\beta\big)$, where β is the bandwidth. These covariance functions typically have a number of hyperparameters which must be adapted. A possible approach is to maximize the marginal likelihood, that is, the likelihood marginalized over the function values, with respect to the hyperparameters $\boldsymbol{\theta}$ of the model. See [12, Chapter 5] for details. The marginal likelihood, also called the evidence of the model, can be used to test Granger causality as follows.

Consider the causality relationship $x \to y$. We propose the following statistic $d_{x\to y}$ that compares the log-evidences of Models 1 and 2

$$d_{x\to y} = \max_{\boldsymbol{\theta}_2} \log P_2(\mathbf{f}_y|X, Y) - \max_{\boldsymbol{\theta}_1} \log P_1(\mathbf{f}_y|Y) \qquad (10)$$

where

$$\begin{aligned} \log P_2(\mathbf{f}_y|X, Y) = &-\frac{1}{2}\mathbf{f}_y^\top (K^2_{zz} + \sigma_y^2 I)^{-1}\mathbf{f}_y \\ &- \frac{1}{2}\log\big|K^2_{zz} + \sigma_y^2 I\big| - \frac{N_{\mathcal{D}}}{2}\log 2\pi \\ \log P_1(\mathbf{f}_y|Y) = &-\frac{1}{2}\mathbf{f}_y^\top (K^1_{yy} + \sigma_y^2 I)^{-1}\mathbf{f}_y \\ &- \frac{1}{2}\log\big|K^1_{yy} + \sigma_y^2 I\big| - \frac{N_{\mathcal{D}}}{2}\log 2\pi \end{aligned}$$

If using a radially Gaussian covariance function for both models, the parameters are $\boldsymbol{\theta}_1 = (\beta_1, \sigma_y^2)$ and $\boldsymbol{\theta}_2 = (\beta_2, \sigma_y^2)$, where β_i is the kernel bandwidth for Model i. As a conclusion, we infer that x causes y if $d_{x\to y} > 0$ because the evidence of Model 2 is then larger than the evidence of Model 1. Causality $y \to x$ can be tested similarly by using the statistic $d_{y\to x}$, defined as above by exchanging the role of y's and x's.

4. ILLUSTRATIONS

4.1. A bivariate example

We consider the coupling of Glass-Mackey like models. The equations are given in discrete time. The coefficients have been chosen to insure stability of the system. Note that increasing the noise variance may cause the system to become unstable. The coupled models are defined by

$$\begin{aligned} x_t &= x_{t-1} - 0.4\left(x_{t-1} - \frac{2x_{t-4}}{1 + x^{10}_{t-4}}\right)y_{t-5} + 0.3y_{t-3} + \varepsilon_{x,t} \\ y_t &= 0.6y_{t-1} + \frac{0.8y_{t-2}}{1 + y^{10}_{t-2}} + \alpha x_{t-2} + \varepsilon_{y,t} \end{aligned}$$

where $\varepsilon_{x,t}$ and $\varepsilon_{y,t}$ are i.i.d. zero-mean Gaussian noises of variance 10^{-2}. It clearly appears that according to Granger's definition of causality, x is caused by y, and y is caused by x if $\alpha > 0$. A training set and a test set of cardinalities $N_{\mathcal{D}} = 500$ and $N_{\mathcal{T}} = 524$ were used. The values of α were successively set to 0, 0.01, 0.1 and 0.2. The lengths $M_{x,y}$ of input vectors were set to 6. The statistics $d_{x\to y}$ and $d_{y\to x}$ are represented in Fig. 1 (left). This experiment clearly confirms the expected results. In particular, as soon as α takes sufficiently large values, the proposed approach correctly detects the coupling between time series x_t and y_t.

4.2. A multivariate example

As previously mentioned, side information must be taken into account when possible to prevent unsound conclusions. It can

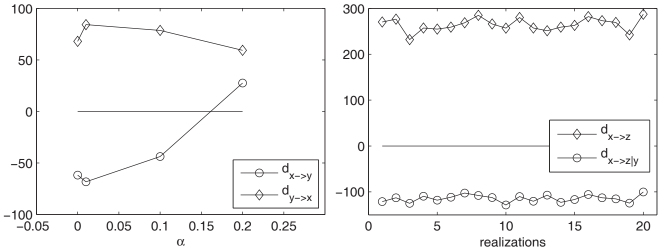

Fig. 1. **Left:** Statistics $d_{x\to y}$ and $d_{y\to x}$ as a function of the coupling strength (bivariate example). **Right:** Statistics $d_{x\to z}$ and $d_{x\to z|y}$ for 20 independent realizations (multivariate example)

be handled within our framework by expanding the input vectors in the learning and test sets accordingly. To illustrate this point, we apply our method to the chain $x \to y \to z$ suggested in [14] and defined by

$$\begin{aligned} x_t &= 1 - ax_{t-1}^2 + \varepsilon_{x,t} \\ y_t &= 0.8(1 - ay_{t-1}^2) + 0.2(1 - ax_{t-1}^2) + \varepsilon_{y,t} \\ z_t &= 0.8(1 - az_{t-1}^2) + 0.2(1 - ay_{t-1}^2) + \varepsilon_{z,t} \end{aligned}$$

with $a = 1.8$and where $\varepsilon_{x,y,z,t}$ are i.i.d. zero-mean Gaussian noises of variance 10^{-4}. The lengths $M_{x,y,z}$ of input vectors were set to 2. The cardinality of the training and test sets was fixed to $N_{\mathcal{D}} = 500$ and $N_{\mathcal{T}} = 524$. The values of statistics $d_{x\to z|y}$ and $d_{x\to z}$ are plotted in Fig. 1 (right) for 20 independent realizations. The bivariate statistic $d_{x\to z}$ which does not take the side information (here contained in the time series y) into consideration, indicates that x causes z. However, when incorporating this side information, the conditional statistic $d_{x\to z|y}$ indicates that x does not cause z conditionally to y. This result is in accordance with the chain $x \to y \to z$.

5. DISCUSSION

This paper presents an original contribution on the use of Gaussian process framework for Granger causality testing. We proposed to use the evidence of the models to design a test of causality. We illustrated the usefulness of our approach with some simulations.

Perspectives of our work include the statistical analysis of our causality test and a full Bayesian implementation. For practical use, in neuroscience in particular, the design of online procedures appears as an interesting opportunity. This however requires some sophisticated sparsification techniques to limit the increase in the dimensions of Gram matrices, as data is recorded. The approach developed in [13] should offer a natural solution to solve this problem. Furthermore, within this context, instantaneous causality is an important concept in practice that has to be handled theoretically.

6. REFERENCES

[1] C. W. J. Granger, "Some recent developments in a concept of causality," *J. of Econometrics*, vol. 39, pp. 199–211, 1988.

[2] J. Geweke, "Measurement of linear dependence and feedback between multiple time series," *J. of the Amer. Stat. Asso.*, vol. 77, pp. 304–313, 1982.

[3] J. Geweke, "Measures of conditional linear dependence and feedback between times series," *J. of the Amer. Stat. Asso.*, vol. 79, no. 388, pp. 907–915, Dec. 1984.

[4] C. Gouriéroux, A. Monfort, and E. Renault, "Kullback causality measures," *Annals of Economics and Statistics*, no. 6-7, pp. 369–410, 1987.

[5] T. J. Mosedale, D. B. Stephenson, M. Collins, and T. C. Mills, "Granger causality of coupled climate processes: Ocean feedback on the north Atlantic oscillation," *Journal of Climate*, vol. 19, pp. 1182–1194, 2006.

[6] B. Gourévitch, R. Le Bouquin-Jeannès, and G. Faucon, "Linear and nonlinear causality between signals: methods, example and neurophysiological applications," *Biol. Cyber.*, vol. 95, no. 4, pp. 349–369, 2006.

[7] J. Rissanen and M. Wax, "Measures of mutual and causal dependence between two time series," *IEEE Trans. on Information Theory*, vol. 33, pp. 598–601, 1987.

[8] S. Kim and E. N. Brown, "A general statistical framework for assessing Granger causality," in *Proc. IEEE ICASSP*, 2010, pp. 2222–2225.

[9] T. Schreiber, "Measuring information transfer," *Phys. Rev. Lett.*, vol. 85, no. 2, pp. 461–465, 2000.

[10] P. O. Amblard and O. J. J. Michel, "On directed information theory and granger causality graphs," *Journal of Computational Neuroscience*, vol. 30, no. 1, pp. 7–16, january 2011.

[11] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hastopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings," *Journal of Computational Neuroscience*, vol. 30, pp. 17–44, 2011.

[12] C. E. Rassmussen and C. K .I. Williams, *Gaussian processes for machine learning*, MIT Press, Cambridge, Ma, USA, 2006.

[13] C. Richard, J. C. M. Bermudez, and P. Honeine, "Online prediction of time series data with kernels," *IEEE Trans. on Signal Processing*, vol. 57, no. 3, pp. 1058–1067, Mar 2009.

[14] D. Marinazzo, M. Pellicoro, and S. Stramaglia, "Kernel-Granger causality and the analysis of dynamical networks," *Phys. Rev. E*, vol. 77, pp. 056215, 2008.