2012 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 23–26, 2012, SANTANDER, SPAIN

KERNELIZING GEWEKE'S MEASURES OF GRANGER CAUSALITY

P. O. Amblard$^{(1,2)}$, *R. Vincent*$^{(1,2)}$

$^{(1)}$ Dept. of Math&Stat
The University of Melbourne, Australia
$^{(2)}$GIPSA-lab, UMR CNRS 5083
Grenoble, France

O. J. J. Michel$^{(2)}$, *C. Richard*$^{(3)}$

$^{(3)}$ Université de Nice Sophia-Antipolis, France
Institut Universitaire de France

ABSTRACT

In this paper we extend Geweke's approach of Granger causality by deriving a nonlinear framework based on functional regression in reproducing kernel Hilbert spaces (RKHS). After giving the definitions of dynamical and instantaneous causality in the Granger sense, we review Geweke's measures. These measures quantify improvement in predicting a time series when the past of another one is taken into account. Geweke's measures are based on linear prediction, and we present an alternative using nonlinear prediction implemented using regularized regression in RKHS. We develop the approach and describe the cross-validation step implemented to optimize the hyperparameters (kernel and regularization parameters). We illustrate the approach on two examples. The first one shows the importance of taking into account side information and possible nonlinear effects. The second one is an illustration of the complete inference problem: surrogate data are generated to create the null hypothesis and the nonlinear measures of causal influence are presented in a test framework.

Index Terms— regression, reproducing kernel Hilbert space, Granger causality

1. INTRODUCTION

When looking for a flow of information between time series, an elegant approach is to consider Granger causality between the time series of interest. Granger causality relies on prediction. A first signal causes a second signal if it helps in the prediction of the second signal. Granger causality was developed essentially in the econometrics community after an early idea of N. Wiener [21, 10]. In the last 10 to 15 years, an increase of interest has been seen in neuroscience especially, but also in climatology ([8, 15]).

The definitions widely accepted rely on conditional independence. Let x_t, y_t, z_t be three time series. The collection of the samples of a signal w up to time t is written w^t. We have the following definitions [10, 4].

Definition 1: Dynamical causality: y does not (dynamically) cause x relatively to (x, y, z) if for all $t \in \mathbb{Z}$,

$$P\big(x_t \big| x^{t-1}, y^{t-1}, z^{t-1}\big) = P\big(x_t \big| x^{t-1}, z^{t-1}\big)$$

Definition 2: Instantaneous coupling: y does not instantaneously cause x relatively to (x, y, z) if for all $k \in \mathbb{Z}$,

$$P\big(x_t \big| x^{t-1}, y^t, z^t\big) = P\big(x_t \big| x^{t-1}, y^{t-1}, z^t\big)$$

In these equalities, $P(.|.)$ stands for the conditional probability distribution (or p.d.f. when they exist).

Dynamical Granger causality states that y causes x if the prediction of x from its past is improved when also considering the past of y. Moreover, dynamical causality is relative to the observation ((x, y, z) in the definitions). Therefore, in testing causality between x and y, taking into account the side information z up to time $t-1$ is fundamental.

Instantaneous coupling quantifies the inclusion of the present of y when estimating the present of x from its past and the past of y. Like dynamical causality, the notion is relative to the whole observation. Note also in this case that the side information is included up to the present time instant. A discussion of the importance of the time horizon chosen for the side information is proposed below.

Testing causality can be done using information theoretic tools, and doing so reveals a very close link between Granger causality and directed information theory [9, 18, 1, 16]. However, estimating information theoretic measures is rather difficult and requires large amount of data. An alternative to information theoretic measures relies on reproducing kernel Hilbert spaces (RKHS) based measures of conditional independence [5].

Here, we use the prediction interpretation and explicitly quantifies the advantage of considering one time series for the prediction of another one. This approach is the original one, and in the Gaussian linear case, it was completely formalized by Geweke in the early 80's.

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

Geweke's measures being linear conditional measures to assess Granger causality, it has appeared natural to design their nonlinear counterparts. Some attempts have been made in this direction over the past, using parametric nonlinear models (see [8] for example). In [14] a kernel based approach was presented, much like the approach we are proposing here. However, if Geweke's measures are based on optimal linear prediction, optimization of the kernel parameters is not made in [14]. In [2], we proposed a Gaussian Process Regression approach which led to causality measures inherently based on nonlinear regression using kernels. It did not however deal with the problem of instantaneous causality.

In this paper, we use regularized regression in RKHS in order to perform a nonlinear functional prediction of the time series. The regression is optimized for a given kernel form using ten fold random cross-validation. When looking for a possible causality from y to x, prediction of x_t is performed twice, doing it with its past only, and with the past of y as well. An index of causality is then derived from the prediction mean square error. The procedure is the same to define an instantaneous causality index.

The paper is organized as follows. In the following section, we briefly recall Geweke's indexes of causality. The prediction procedure is developed in section 3.1 and is applied in 3.2 to define the kernel version of Geweke's measures. The practical set-up, including cross-validation and testing procedures, is developed in section 4. Section 5 then provides two illustrations. A chain of nonlinear systems shows the importance of side information and of nonlinear processing. A complete four dimensional example illustrates the whole inference problem. We give some future directions in the conclusion.

2. GEWEKE'S MEASURES

We briefly recall Geweke's measures, introducing his two important papers [6, 7]. In the 1982 paper, the measures of causality are evaluated between two time series (possibly multivariate) whereas the 1984 paper introduces side information in the measures. The 1982 framework is thus included in the 1984 framework, and therefore not detailed here. We recover it by suppressing side information. Furthermore, for the sake of simplicity, we concentrate on the causality between two univariate time series x_t, y_t when a (possibly multivariate) third time series z_t is considered as side information. At time t, let x^{t-1} be the collection of the past samples of signal x. Geweke's measures rely on linear prediction and are well interpreted in the model

$$\left\{ \begin{array}{lcl} x_t &=& \ell_x(x^{t-1}) + \ell_{yx}(y^{t-1}) + \ell_{zx}(z^{t-1}) + \varepsilon_{x,t} \\ y_t &=& \ell_{xy}(x^{t-1}) + \ell_y(y^{t-1}) + \ell_{zy}(z^{t-1}) + \varepsilon_{y,t} \end{array} \right. \quad (1)$$

where the ℓ's denote linear functionals. For example, $\ell_{yx}(y^{t-1}) = \sum_{n\geq 1} h_n y_{t-n}$ for some $\{h_n\}$. The residuals $\varepsilon_{.,t}$ are assumed to be sequences of i.i.d. zero mean Gaussian random variables. These two sequences can be correlated. We suppose that, if they are correlated, correlation is instantaneous. Otherwises the linear functionals are modified to take the delay into account.

When estimating x_t linearly from x^n, y^m, z^o, where indices n, m, o are either t or $t-1$, let $\sigma^2_\infty(x_t|x^n, y^m, z^o) = \lim_{t\to+\infty} \sigma^2(x_t|x^n, y^m, z^o)$ be the asymptotic minimum mean square error. The abuse of notation allows to keep track of the horizon used in the prediction. Geweke's indices measure the gain in predicting x_t by incorporating other variables than its past as regressors. These indices are:

$$F_{y\to x\|z} = \log \frac{\sigma^2_\infty(x_t|x^{t-1}, z^{t-1})}{\sigma^2_\infty(x_t|x^{t-1}, y^{t-1}, z^{t-1})} \quad (2)$$

$$F_{x.y\|z} = \log \frac{\sigma^2_\infty(x_t|x^{t-1}, y^{t-1}, z^{t})}{\sigma^2_\infty(x_t|x^{t-1}, y^{t}, z^{t})} \quad (3)$$

$F_{x\to y\|z}$ is defined in the same way, and it can be shown that $F_{x.y\|z}$ is symmetrical in x and y. $F_{y\to x\|z}$ measures the advantage of including the past of y when predicting x from its past and the past of the side information. If there is no advantage, the measure is zero, whereas if there is an advantage, the measure is strictly positive. In this case, y is said to be a *prima facie* cause of x, relatively to x, y, z.

3. KERNELIZED GEWEKE'S MEASURE

3.1. Prediction with kernels

Geweke searched for the best predictor in the class of linear functionals. Other approaches were derived in parametric classes of nonlinear functionals ([18]). Here, we seek the best predictor in an *a priori* infinite dimensional Hilbert space of functions from the input space to $\mathbb{R}$. The prediction of signal x can be written generically as $x_t = f(w^{t-1})$ where, for the application developed here, w may be either x alone, x and z or x, y and z. We suppose that $x_t \in \mathbb{R}$ and that $w^{t-1} \in \mathbb{X}$, where $\mathbb{X}$ is typically $\mathbb{R}^p$. Here, p is linked to the memory considered for predicition. Considering the M past samples, p will be equal to M if $w = x$, equal to $2M$ if $w = (x, z)$ and to $3M$ if $w = (x, y, z)$. Then we consider a symmetric positive definite function $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ and the unique reproducing kernel Hilbert space $\mathcal{H}$ of functions from $\mathbb{X}$ to $\mathbb{R}$ associated to it [19, 20]. We now seek the best predictor in this reproducing kernel Hilbert space, optimal in the MMSE sense.

From a practical point of view, we have to learn the best predictor from a set of data $(x_t, w^{t-1})_i \in \mathbb{R} \times \mathbb{X}$ where $i = 1, \dots, N$. The problem is then to solve

$$f^\star = \arg\min_{f\in\mathcal{H}} \sum_i |x_{t,i} - f(w_i^{t-1})|^2 + \lambda \|f\|^2_{\mathcal{H}} \quad (4)$$

where we have added the regularization term to avoid overfitting ($\lambda\|.\|^2_{\mathcal{H}}$ stands for the norm in $\mathcal{H}$). This problem is the problem of regression and its solution is well-known [19]. The representer theorem implies that function f belongs to the finite dimensional subspace of $\mathcal{H}$ generated by $k(., w_i^{t-1}), i = 1, \ldots, N$, and therefore writes $f(.) = \boldsymbol{\alpha}^\top \boldsymbol{k}_{w^{t-1}}(.)$ where $\alpha \in \mathbb{R}^N$ and $\boldsymbol{k}_w(.)$ is an N dimensional vector of functions with entries $(\boldsymbol{k}_w(.))_i := k(., w_i^{t-1})$. Problem (4) then reduces to an optimization problem over $\boldsymbol{\alpha}$. Let $\boldsymbol{K}_w$ be the Gram matrix with entries $(\boldsymbol{K}_w)_{ij} := k(w_i^{t-1}, w_j^{t-1})$, and let $\boldsymbol{I}$ stand for the identity matrix. Denote as $\boldsymbol{x}_t$ the N dimensional vector with entries $(\boldsymbol{x}_t)_i := x_{t,i}$. Then the optimal parameter vector $\boldsymbol{\alpha}^\star$ is given by

$$\boldsymbol{\alpha}^\star = (\boldsymbol{K}_w + \lambda \boldsymbol{I})^{-1} \boldsymbol{x}_t \tag{5}$$

and the optimal prediction based on an observation w^{t-1} can be expressed as

$$\widehat{x_t} = f^\star(w^{t-1}) = \boldsymbol{k}_w(w^{t-1})^\top (\boldsymbol{K}_w + \lambda \boldsymbol{I})^{-1} \boldsymbol{x}_t \tag{6}$$

We can evaluate the mean square prediction error by averaging over T realizations by

$$\sigma^2(x_t | w^{t-1}) = \frac{1}{T} \sum_{j=1}^{T} |x_{t,j} - f^\star(w_j^{t-1})|^2 \tag{7}$$

Note at this point that we have obtained the best predictor in the RKHS generated by k for a particular value of λ. However and as usual, k depends on some parameters. For example, if we consider the Gaussian kernel $k(x, y) = \exp(-\|x - y\|^2/\beta^2)$, the solution is optimal for the fixed pair of parameters (β, λ). We do not guarantee that for another pair the solution cannot be better. Therefore, we also have to optimize with respect to these parameters in order to obtain the best solution for the particular form of the objective and of the kernel. This optimization would be seen as marginalization in a Bayesian point of view. As known however, this optimization cannot be done in closed form, and we do it using cross-validation. The whole procedure will be explained in section 4.1. From now on, we suppose that the optimization over these parameters has been done, and that we have obtained the optimal predictor for a given form of the kernel and the regularized least-square objective function.

3.2. Kernel Geweke measures

We use the results of the previous section to directly generalize Geweke's linear measures of causality to the kernel framework adopted here. The bivariate kernel Geweke's measures are defined as

$$G_{y \to x} := \log \frac{\sigma^2(x_t | x^{t-1})}{\sigma^2(x_t | x^{t-1}, y^{t-1})} \tag{8}$$

$$G_{x.y} := \log \frac{\sigma^2(x_t | x^{t-1}, y^{t-1})}{\sigma^2(x_t | x^{t-1}, y^t)} \tag{9}$$

and there multivariate counterparts read

$$G_{y \to x \| z} := \log \frac{\sigma^2(x_t | x^{t-1}, z^{t-1})}{\sigma^2(x_t | x^{t-1}, y^{t-1}, z^{t-1})} \tag{10}$$

$$G_{x.y \| z} := \log \frac{\sigma^2(x_t | x^{t-1}, y^{t-1}, z^t)}{\sigma^2(x_t | x^{t-1}, y^t, z^t)} \tag{11}$$

Theoretically, these indexes are greater than or equal to zero, strict positivity indicating causality. As in the linear case, dynamical causality assessed by strict positivity of $G_{y \to x}$ relies on the improvement in the predicition of x_t provided by the past of y_t. Likewise, $G_{x.y}$ and $G_{x.y\|z}$ measure instantaneous coupling as the improvement provided by the present of y_t in the estimation of x_t given its past and the past of y_t.

Note the time horizon chosen for the side information in the definition of the causality measures. For the dynamical causality index $G_{y \to x \| z}$, z is considered up to $t-1$. If considered up to t, instantaneous coupling would come into play and lead to erroneous conclusion regarding dynamical causality. In $G_{x.y\|z}$, the time horizon for side information is t. The choice $t-1$ is possible. However, the graph induced would be an independence graph instead of a conditional independence graph which is the usual framework adopted in graphical modeling, especially for Markov properties of the graph. For more details on this discussion, see [3].

4. DETAILS ON THE PRACTICAL SET-UP

4.1. Cross-validation

For the optimization of λ and of the kernel parameters, we use a randomized 10-fold cross-validation procedure, as described below.

Suppose we observe the three time series x_t, y_t, z_t for $t = 1, \ldots, n_l$ and we want to calculate $G_{y \to x \| z}$. We choose a time horizon p, and we create the learning set $(x_{t,i}, x_i^{t-1}, y_i^{t-1}, z_i^{t-1})$ for $i = 1, \ldots, n_l - p$, where $x_{t,i} = x_i$ and $w_i^{t-1} = (w_{i-1}, \ldots, w_{i-p})$ for $w = x, y, z$. The learning set is split into ten subsets of equal size in which every element has been randomly picked from the original learning set, once and only once. Each subset is used once for the validation step, while the learning step is processed over the remaining nine other subsets. This procedure is repeated ten times so that every subset is used once and only once for validation. The mean square error is obtained as the average of the ten mean square errors calculated on every validation subset. This is repeated over a grid in λ and in the parameters of the kernel, and this allows to find the optimal parameters (the grid search is implemented using five successive logarithmic subdivisions of an initial set). We finally evaluate again the mean square error of prediction on the whole learning set with the optimal parameters.

4.2. Testing

As mentionned in [7], even in the linear Gaussian case, the multivariate measures of causality have no explicitly known distribution. Therefore, the evaluation of the errors and power of the test associated to these measures is impossible from a theoretical perspective. A numerical approach is needed. Here, we use random permutations of some data to simulate the null hypothesis of no causality.

For example in the test

$$\left\{ \begin{array}{lcll} H_0 & : & G_{y\to x\|z} & = & 0 \\ H_1 & : & G_{y\to x\|z} & > & 0 \end{array} \right. \tag{12}$$

we create artificially H_0 by shuffling the data which are supposed to be responsible for H_1, *i.e.* in the example considered, we shuffle the vectors $\{y_i^{t-1}\}$. Precisely, let π be a random permutation of the first $n_l - p$ integers, we generate H_0 by considering $y_{\pi(i)}^{t-1}$ instead of y_i^{t-1} in the learning set. For n_r random independent random permutations we evaluate $G^0_{y\to x\|z}$. We can then estimate from this n_r experiments the probability of error of the first kind (or false alarm probability) $P(\eta) = \Pr(G_{y\to x\|z} > \eta \big| H_0)$ by $\sum_{n=1}^{n_r} \mathbf{1}(G^0_{y\to x\|z}(n) > \eta)/n_r$, $\mathbf{1}(A)$ being the characteristic function of the set A. This in particular allows to determine the threshold η for which the error of the first kind is bounded by say α. Note that for this we do not explicitly estimate the probability but rather use the estimation of the cumulative probability distribution function by sorting the n_r numbers $G^0_{y\to x\|z}(n)$.

5. ILLUSTRATIONS

We provide two illustrations that we have already studied from an information theory point of view [3]. The first one highlights the joint necessity of nonlinear processing and accounting for side information. The second illustration provides a full analysis of a four dimensional example.

5.1. A chain

For this example, let $\varepsilon_{x,y,z,t}$ be three independent zero mean, unit variance white Gaussian noise, and consider the following model

$$\left\{ \begin{array}{lcl} x_t & = & ax_{t-1} + \varepsilon_{x,t} \\ y_t & = & by_{t-1} + d_{xy}x^2_{t-1} + \varepsilon_{y,t} \\ z_t & = & cz_{t-1} + c_{yz}y_{t-1} + \varepsilon_{z,t} \end{array} \right. \tag{13}$$

where $a = 0.2, b = 0.5, c = 0.8, d_{xy} = 0.8, c_{yz} = 0.7$. Note the quadratic coupling from x to y. As seen in the definition, the chain $x \to y \to z$ is a Markov chain, and therefore we expect that x causes z relatively to (x, z), but x does not cause z relatively to (x, y, z). We have generated $n_r = 500$ realizations of the model, each on $n_l = 500$ samples. On each realization, we evaluate the bivariate index $G_{x\to z}$ and the multivariate index $G_{x\to z\|y}$. The calculation is performed as described in the previous sections with $p = 2$. In this example we have used the linear kernel $k(x, y) = x^\top y$ (panel (a) in figure 1) and the Gaussian kernel (panel (b)). Note that using the linear kernel leads to the usual linear Geweke's measures.

We display here the results in the form of histograms of the G's evaluated on 20 bins. The plots are depicted in figure (1). From panel (a) we conclude that the linear measures predict non causality from x to z, and this relatively either to (x, z) or to (x, y, z). This erroneous conclusion is due to the quadratic term in the link from x to y. Using the Gaussian kernel leads to the expected conclusion. In the bivariate analysis, we conclude that x causes z since the histogram of $G_{x\to z}$ is clearly centered to a strictly positive number with a narrow support. Taking the side information y into account, the measure dramatically decreases, and $G_{x\to z\|y}$ is now centered around zero with a very small dispersion. Therefore we conclude from panel (b) that the existing causality from x to z is mediated by y, as the model suggests.

Although the measures should be strictly positive, note that some negative values are observed in practice. This is due to estimation residuals.

5.2. A four-dimensional example

In this example, let $\varepsilon_{w,x,y,z,t}$ be an i.i.d. sequence of zero mean four dimensional Gaussian vectors, with covariance

$$\Gamma_\varepsilon = \begin{pmatrix} 1 & \rho_1 & 0 & \rho_1\rho_2 \\ \rho_1 & 1 & 0 & \rho_2 \\ 0 & 0 & 1 & \rho_3 \\ \rho_1\rho_2 & \rho_2 & \rho_3 & 1 \end{pmatrix} \tag{14}$$

where we set $\rho_1 = 0.66, \rho_2 = 0.55$ and $\rho_3 = 0.48$. Now consider the following model

$$\left\{ \begin{array}{lcl} w_t & = & 0.2w_{t-1} + 0.3z_{t-1} - 0.2ex^2_{t-1} + \varepsilon_{w,t} \\ x_t & = & 0.3x_{t-1} + 0.3z^2_{t-1} + \varepsilon_{x,t} \\ y_t & = & -0.8y_{t-1} + 0.8x_{t-1} - 0.5x^2_{t-1} + \varepsilon_{y,t} \\ z_t & = & -0.4z_{t-1} + 0.2w_{t-1} + \varepsilon_{z,t} \end{array} \right. \tag{15}$$

The model may be synthesized by what is called a causality graph [4]. This is a mixed graph, whose vertices represent the signals, a directed edge from x to y represents dynamical causality from x to y relatively to the four signals, and an undirected edge between x to y represents instantaneous coupling between x and y relatively to the four signals. The graph and the associated adjacency matrices are depicted in figure (2). Note that the undirected graph associated to instantaneous coupling corresponds, according to def. 2, to the conditional independence graph of the noise. The absence of an edge in this graph corresponds to a zero in the precision matrix (inverse of the covariance matrix) in the entry corresponding to the signals tested.

We have generated $n_r = 500$ realizations of the four times series, each of length 660 samples. Learning was performed

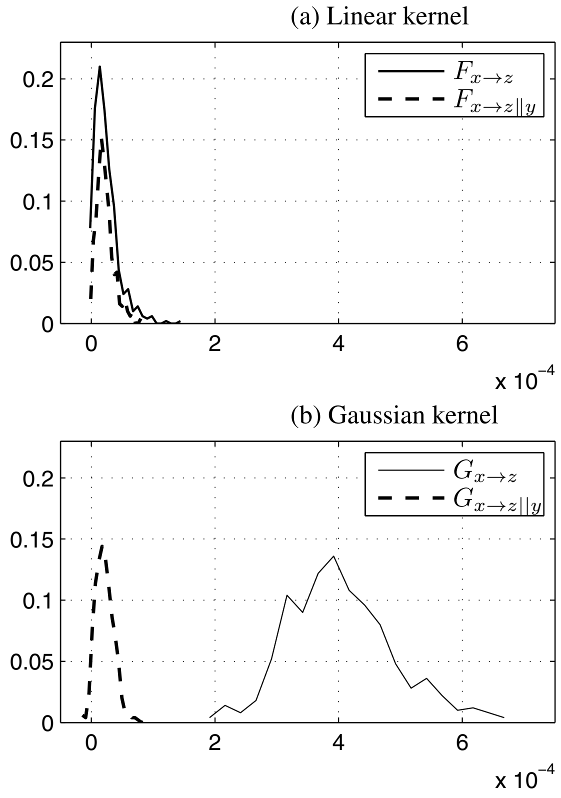
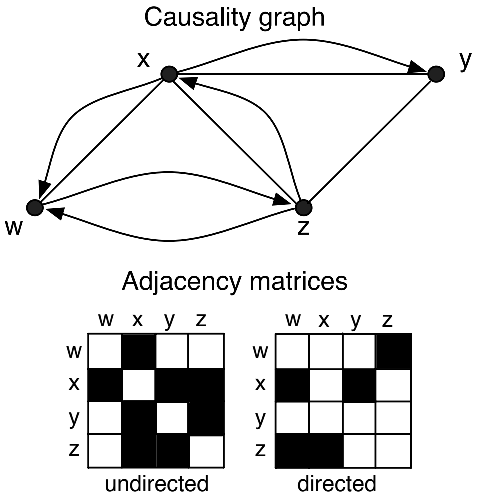


Fig. 1. Histograms of dynamical causality indexes $x \to z$ in a chain of nonlinear systems $x \to y \to z$. (a) Linear Geweke's measure calculated using the approach developed in the paper with the linear kernel. (b) Nonlinear Geweke's dynamical causality index calculated as described in the paper using a Gaussian kernel.

for each realization using the Gaussian kernel. The time horizon p considered here was chosen to be $p = 3$. Note that this choice should be included in the parameter set, and should be optimized during the cross-validation step. This will be considered in future work. Each measure was calculated on the original realization (denoted $G_{\dots}$) and on the shuffled data mimicking the null hypothesis (measure denoted as $G^0_{\dots}$ in that case). We insist again on the fact that the only shuffled variable is the variable possibly responsible of H_1. For example, in testing instantaneous coupling between x and y (relatively to (w, x, y, z)), the deviation from H_0 is due the the present of y, and this series only is shuffled: in the estimation of $\sigma^2(x_t \big| x^{t-1}, y_t, y^{t-1}, w^t, z^t)$, the only shuffled data are the realizations of y_t (the realizations of y^{t-1} are not shuffled!).

For each possible pair, we estimated the threshold to obtain a familywise probability of false alarm of $\alpha = 10\%$. Therefore to be sure to obtain this rate, we used the Bonferronni correction, a very well-known and very conservative correction [13]. Some better and less conservative correction such as false discovery rate correction (FDR) could also be implemented. Thus practically, the pairwise probability of false alarm is set to $\alpha/12$ for dynamical causality testing and $\alpha/6$ for instantaneous causality testing (because of symmetry). Then, for each pair and each test we have a threshold, and we can effectively test on the the original data. We can then estimate the probability of deciding H_1 from the n_r realizations. We plot the results in figure (3) in the form of matrices of these estimated probabilities. As illustrated in these figures, taking side information into account is necessary to infer the correct graph. For example, in plot (a), the adjacency matrix of the undirected graph inferred is erroneous. However, the structure is recovered in plot (b), up to the false alarm rate. The same conclusion holds for the dynamical causality (see plots (c) and (d)). In (c), the probability of deciding H_1 for the entries (w, x), (x, z), (y, w) and (z, y) largely exceeds the false alarm probability and should lead to the assignment of a directed edge although no such ink exists. Note that the results are obtained on short length of signals (learning done on 660 samples). This example was already considered in [3] where the measures considered are based on directed information theory. In these works, the results are comparable in quality, but are obtained on 5 times longer sample size.



	w	x	y	z
w	□	■	□	□
x	■	□	■	■
y	□	■	□	■
z	□	■	■	□

undirected

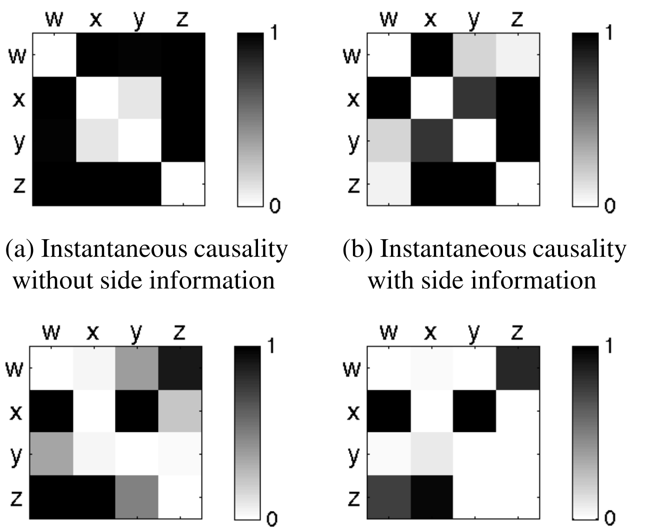
	w	x	y	z
w	□	□	□	■
x	■	□	■	□
y	□	□	□	□
z	■	■	□	□

directed

Fig. 2. Causality graph of the second example developed in the paper. Arrows stand for dynamical causality and lines for instantaneous coupling. We represent also the adjacency matrices associated to the two types of edges in the causality graph.

6. SOME FUTURE WORKS

As already mentioned, including the time horizon as a parameter to be optimized in the cross-validation step is definitely one of the next tasks. Furthermore, we develop iterative version of the approach, having in mind on-line applications and



(a) Instantaneous causality without side information

(b) Instantaneous causality with side information

(c) Dynamical causality without side information

(c) Dynamical causality with side information

Fig. 3. Causal inference in the 4 dimensional model presented in the text. The matrices represent the probability that a causality measure exceeds the threshold defined to provide a familywise false alarm probability of 10%. The null hypothesis is created using random permutations of the variable responsible for the alternative, allowing empirical evaluation of the threshold.

nonstationary analysis. This is of utmost importance for *e.g.* neuroscience or climatology applications.

7. REFERENCES

[1] 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.

[2] P.-O. Amblard, O. J. J. Michel, C.Richard, and P. Honeine. A Gaussian process regression approach for testing Granger causality between time series data. In *proc. ICASSP, Osaka, Japan*, 2012.

[3] P. O. Amblard and O. J. J. Michel. Causal conditioning and instantaneous coupling in causality graphs. *submitted to Signal Processing*, arXiv:1203.5572, 2012.

[4] M. Eichler. Graphical modelling of multivariate time series. *Proba. Theory Relat. Fields*, DOI 10.1007/s00440-011-0345-8, 2011.

[5] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *NIPS*, 2007.

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

[7] 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.

[8] 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.

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

[10] C. W. J. Granger. Testing for causality : a personal viewpoint. *Journal of Economic Dynamics and Control*, 2:329–352, 1980.

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

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

[13] E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses, 3rd ed.* Springer, 2005.

[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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] B. Schölkopf and A. J. Smola. *Learning with kernels*. MIT Press, Cambridge, Ma, USA, 2002.

[20] I. Steinwart and A. Christmann. *Support vector machines*. Springer, 2008.

[21] N. Wiener. The theory of prediction, in *Modern mathematics for the engineer*, E.F. Beckenbach ed., pages 165–190. MacGrawHill, 1956.