Online Nonlinear Granger Causality Detection by Quantized Kernel Least Mean Square*

Hong Ji[1], Badong Chen[1], Zejian Yuan[1], Nanning Zheng[1],
Andreas Keil[2], and Jose C. Príncipe[2]

[1] The Institute of Artificial Intelligence and Robotics,
Xian Jiaotong University, 28 Xianning West Road, Xian 710049, China
[2] Center for the Study of Emotion and Attention,
University of Florida, Gainesville, Florida 32611, USA
{itsjihong,yzejian}@gmail.com
{chenbd,nnzheng}@mail.xjtu.edu.cn
{akeil,principe}@cnel.ufl.edu

Abstract. Identifying causal relations among simultaneously acquired signals is an important challenging task in time series analysis. The original definition of Granger causality was based on linear models, its application to nonlinear systems may not be appropriate. We consider an extension of Granger causality to nonlinear bivariate time series with the universal approximation capacity in reproducing kernel Hilbert space (RKHS) while preserving the conceptual simplicity of the linear model. In particular, we propose a computationally simple online measure by means of quantized kernel least mean square (QKLMS) to capture instantaneous causal relationships.

Keywords: Granger causality, kernel methods, quantized kernel least mean square(QKLMS), nonlinear time series.

1 Introduction

The problem of quantifying causal connectivity among simultaneously acquired time series has received considerable attention in the recent years due to its growing applicability in economy [1], neuroscience [2,3,4], medical and clinical science [5], and many others. One approach to evaluate causal relations between two time series is to examine if one series is better predicted by adding knowledge from the other. This was originally proposed by Wiener [6] and later formalized by Granger in the context of linear regression models of stochastic processes [1]. In particular, if the prediction error of the first time series is reduced by incorporating measurements from the second time series, then the second time series is said to have a causal inference on the first time series. By exchanging the roles of the two time series, the causal influence in the opposite direction can be addressed.

* This work was supported by NSFC grant No. 61372152.

C.K. Loo et al. (Eds.): ICONIP 2014, Part II, LNCS 8835, pp. 68–75, 2014.

As a technique to understand the directed connectivity of the underlying mechanisms, Granger causality has been well explored and construed in many different ways. Also, there is a freely available software toolbox incorporating these methods to facilitate its broadly application in neuroscience data analysis [7]. However, since Granger causality was formulated as linear regression, its application to nonlinear systems, such as brain signal that is highly nonlinear at many levels of description, may not be appropriate. There are several competing approaches to this problem. A simple solution [8] is to fit autoregressive coefficients to Taylor expansions of the data, but this method requires estimating a large number of parameters. Alternative approaches include the radial basis functions (RBFs) [9] and kernel methods. The kernel methods transform the data into a high dimensional reproducing kernel Hilbert space (RKHS) such that appropriate linear methods can be applied on the transformed data[10]. Most of these methods, however, assume the stationarity of the signals.

In this work, we propose a computationally simple online kernel method for causality detection, called the twin quantized kernel least mean square (twin-QKLMS) , which is able to capture the causal relations between nonlinear and non-stationary time series.

2 Granger Causality

2.1 Linear Modeling

Linear Granger causality is defined based on vector autoregressive model [1]. Let $X \equiv \{\overline{x_k}\}_{k=1,...,N}$ and $Y \equiv \{\overline{y_k}\}_{k=1,...,N}$ be two time series of N simultaneously measured quantities. Usually the stationarity of the time series is required. For $i = 1$ to M (where $M = N - m$, m being the order of the model), we denote $x_i = \overline{x_{i+m}}$, $y_i = \overline{y_{i+m}}$, $\mathbf{x}(i) = (\overline{x_{i+m-1}}, \overline{x_{i+m-2}}, \ldots, \overline{x_i})$ and $\mathbf{y}(i) = (\overline{y_{i+m-1}}, \overline{y_{i+m-2}}, \ldots, \overline{y_i})$ and treat these quantities as M realizations of the stochastic variables $(x, y, \mathbf{x}, \mathbf{y})$. The following model is then considered:

$$\begin{aligned} x &= \mathbf{w}_1 \cdot \mathbf{x} + \xi_x \\ y &= \mathbf{w}_2 \cdot \mathbf{y} + \xi_y \end{aligned} \tag{1}$$

Here $\{\mathbf{w}\}$ being m-dimensional real vectors to be estimated from data, ξ_x and ξ_y being the residuals (prediction errors) for each time series when predicted solely based on the knowledge of its own past values. We denote their variance as $\epsilon_x = var(\xi_x)$ and $\epsilon_y = var(\xi_y)$ which are equal to the mean square prediction errors since zero mean has been guaranteed by pre-processing. The temporal dynamics of the two time series can be described by a bivariate autoregressive model:

$$\begin{aligned} x &= \mathbf{w}_{11} \cdot \mathbf{x} + \mathbf{w}_{12} \cdot \mathbf{y} + \xi_{x|Y} \\ y &= \mathbf{w}_{21} \cdot \mathbf{x} + \mathbf{w}_{21} \cdot \mathbf{y} + \xi_{y|X} \end{aligned} \tag{2}$$

Similarly, we define $\epsilon_{x|y} = var(\xi_{x|Y})$ and $\epsilon_{y|x} = var(\xi_{y|X})$. If the prediction of x improves by incorporating the past values of series Y, that $\epsilon_{x|y}$ is smaller than

ϵ_x, then Y has a causal influence on X. Analogously, if $\epsilon_{y|x}$ is smaller than ϵ_y, then X has a causal influence on Y. The magnitude of this interaction can be measured by the log ratio of the prediction error variances:

$$\begin{aligned} F_{Y \to X} &= ln \frac{\epsilon_x}{\epsilon_{x|y}} \\ F_{X \to Y} &= ln \frac{\epsilon_y}{\epsilon_{y|x}} \end{aligned} \tag{3}$$

The maximum of both terms

$$F_{XY} = max\{F_{Y \to X}, F_{X \to Y}\} \tag{4}$$

represents a simple measure for the strength of directional and/or bi-directional interaction.

2.2 Nonlinear Modeling in RKHS

The mapping from input to feature space is induced by a Mercer kernel which is a continuous, symmetric, and positive definite function $\kappa : \mathbb{X} \times \mathbb{X} \to \mathbb{R}, \mathbb{X} \subseteq \mathbb{R}^m$ [12,13]. The Gaussian kernel is widely used for its proved universal approximation property for any continuous function [14].

$$\kappa(\mathbf{x}, \mathbf{x}') = \exp(\frac{-\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}) \tag{5}$$

where $\sigma > 0$ is the kernel size (or kernel bandwidth). According to the Mercer theorem [11,12], any Mercer kernel $\kappa(\mathbf{x}, \mathbf{x}')$ induces a mapping $\boldsymbol{\varphi}$ such that the inner product between the transformed input data (feature vectors) satisfies $\langle \boldsymbol{\varphi}(\mathbf{x}), \boldsymbol{\varphi}(\mathbf{x}') \rangle = \kappa(\mathbf{x}, \mathbf{x}')$.

We now construct the autoregressive model and bivariate model in transformed feature space.

$$\begin{aligned} x &= \boldsymbol{\Omega_1} \cdot \boldsymbol{\varphi}(\mathbf{x}) + \xi_x \\ y &= \boldsymbol{\Omega_2} \cdot \boldsymbol{\varphi}(\mathbf{y}) + \xi_y \end{aligned} \tag{6}$$

with the corresponding prediction error variance ϵ_x and ϵ_y.

$$\begin{aligned} x &= \boldsymbol{\Omega_{11}} \cdot \boldsymbol{\varphi}(\mathbf{x}) + \boldsymbol{\Omega_{12}} \cdot \boldsymbol{\psi}(\mathbf{y}) + \xi_{x|Y} \\ y &= \boldsymbol{\Omega_{21}} \cdot \boldsymbol{\varphi}(\mathbf{y}) + \boldsymbol{\Omega_{22}} \cdot \boldsymbol{\psi}(\mathbf{x}) + \xi_{y|X} \end{aligned} \tag{7}$$

where $\{\boldsymbol{\Omega}\}$ are the weight vectors in feature space (infinite dimensional for the Gaussian kernel case). The prediction errors to minimize are defined as:

$$\begin{aligned} \epsilon_{x|y} &= \frac{1}{M} \sum_{i=1}^{M} [x_i - \boldsymbol{\Omega_{11}} \cdot \boldsymbol{\varphi}(\mathbf{x}(i)) - \boldsymbol{\Omega_{12}} \cdot \boldsymbol{\psi}(\mathbf{y}(i))]^2 \\ \epsilon_{y|x} &= \frac{1}{M} \sum_{i=1}^{M} [y_i - \boldsymbol{\Omega_{21}} \cdot \boldsymbol{\varphi}(\mathbf{y}(i)) - \boldsymbol{\Omega_{22}} \cdot \boldsymbol{\psi}(\mathbf{x}(i))]^2 \end{aligned} \tag{8}$$

By using the kernel trick, we can efficiently compute the inner product output by kernel evaluation without knowing the exact form of mapping. In the following we take the bivariate model to predict x as an example, the prediction of y can be derived analogously:

$$\begin{aligned} f(\mathbf{x}) = \boldsymbol{\Omega}_{\mathbf{11}} \cdot \boldsymbol{\varphi}(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i \boldsymbol{\varphi}(\boldsymbol{c}_i)^T \boldsymbol{\varphi}(\mathbf{x}) = \sum_{p=1}^{l} \alpha_i \kappa(\boldsymbol{c}_i, \mathbf{x}) \\ g(\mathbf{y}) = \boldsymbol{\Omega}_{\mathbf{12}} \cdot \boldsymbol{\varphi}(\mathbf{y}) = \sum_{i=1}^{s} \beta_i \boldsymbol{\varphi}(\boldsymbol{c}_i^{'})^T \boldsymbol{\varphi}(\mathbf{y}) = \sum_{i=1}^{l} \beta_i \kappa(\boldsymbol{c}_i^{'}, \mathbf{y}) \end{aligned} \tag{9}$$

where $\{\boldsymbol{c}_i, \alpha_i\}_{i=1}^{l}$ and $\{\boldsymbol{c}_i^{'}, \beta_i\}_{i=1}^{s}$ are the parameters to learn. Note we assume that x is the sum of a term depending solely on $\mathbf{x}$ and a term depending solely on $\mathbf{y}$ instead of the general bivariate model which are depending on the appending vector $(\mathbf{x}\ \mathbf{y})$:

$$x = f(\mathbf{x}) + g(\mathbf{y}) \tag{10}$$

It has been proposed that any prediction scheme providing a nonlinear extension of Granger causality should satisfy the following (P1) property [9]: *if* $\mathbf{y}$ *is statistically independent of* $\mathbf{x}$ *and* x*, then* $\epsilon_x = \epsilon_{x|y}$*; if* $\mathbf{x}$ *is statistically independent of* $\mathbf{y}$ *and* y*, then* $\epsilon_y = \epsilon_{y|x}$*;* Let us suppose $\mathbf{y}$ is statistically independent of $\mathbf{x}$ and x. Then $\boldsymbol{\varphi}(\mathbf{x})$ is uncorrelated with x and with $\boldsymbol{\psi}(\mathbf{y})$. It follows that

$$\begin{aligned} \epsilon_{x|y} =& var[x - \boldsymbol{\Omega}_{\mathbf{11}} \cdot \boldsymbol{\varphi}(\mathbf{x}) - \boldsymbol{\Omega}_{\mathbf{12}} \cdot \boldsymbol{\psi}(\mathbf{y})] \\ =& var[x - \boldsymbol{\Omega}_{\mathbf{11}} \cdot \boldsymbol{\varphi}(\mathbf{x})] + var[\boldsymbol{\Omega}_{\mathbf{12}} \cdot \boldsymbol{\psi}(\mathbf{y})] \end{aligned} \tag{11}$$

To minimize $\epsilon_{x|y}$ it follows that $\boldsymbol{\Omega}_{\mathbf{12}} = 0$ which satisfy P1 property.

2.3 Twin-QKLMS Causality Detector

The QKLMS is one of the most simple and efficient online kernel adaptive filter algorithm. It natrally creates a growing radial-basis function network, learning network topology adaptively. It has been verified that a sufficient condition for mean square convergence and a bounded theoretical value of the steady-state excess mean square error [11]. Inheriting from KLMS, it does not need explicit regularization to obtain solutions that generalize appropriately [15]. In this section we propose a novel online detector for causality analysis and employ a twin QKLMS to the bivariate nonlinear model described in previous section. We name the model as "twin-QKLMS" to emphasize that two QKLMS filters work in parallel in an online mode. The reason to choose QKLMS comes from the key properties highlight below:

- QKLMS uses quantization to compress the input space so as to constrain the network size growth. Different from conventional sparsification methods normally discarding samples simply, the "redundant" data are used to locally update the coefficient of the closest center, which help to achieve better accuracy and a more compact network.

- The codebook is trained directly from online samples and is adaptively growing, unlike RBF constrains the network size to a fixed size (by cluster to n prototypes).

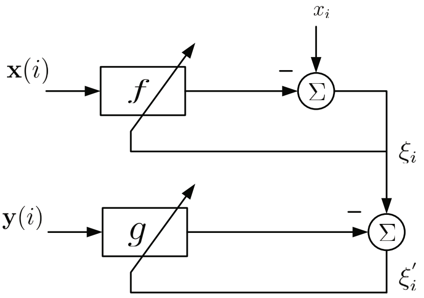


Fig. 1. Twin-QKLMS model as an estimator of mapping $x = f(\mathbf{x}) + g(\mathbf{y})$, filter $g(\cdot)$ is employed to predict the residual of $x - f(\mathbf{x})$ incorporating the knowledge from input $\boldsymbol{y}$

The twin-QKLMS models a dual filtering structure in Fig.1. Denote $\varphi(i) = \varphi(\boldsymbol{u}(i))$, quantizing the input vector in the filter update equation. Using the basics of QKLMS, we propose the twin-QKLMS algorithm below.

$$\begin{cases} f_0 = 0 \\ g_0 = 0 \\ \xi_i = x_i - f_{i-1}(\mathbf{x}(i)) \\ f_i = f_{i-1} + \lambda \xi_i \kappa(\boldsymbol{Q}[\mathbf{x}(i)], \cdot) \\ \xi_i^{'} = \xi_i - g_{i-1}(\mathbf{y}(i)) \\ g_i = g_{i-1} + \eta \xi_i^{'} \kappa(\boldsymbol{Q}[\mathbf{y}(i)], \cdot) \end{cases} \tag{12}$$

where ξ_i is the prediction error at iteration i by predicting x_i with filter f_{i-1} enclosed input $\boldsymbol{x}(i)$, $\xi_i^{'}$ is the prediction error that predict residual ξ_i with filter g_{i-1} that incorporate the knowledge of the other series $\boldsymbol{y}(i)$, λ and η are the step size, $\boldsymbol{Q}[.]$ denotes the quantization operator. f_i is the composition of $\boldsymbol{\Omega_f}$ and φ, that is $f_i = \boldsymbol{\Omega_{11}(i)^T}\boldsymbol{\varphi(\cdot)}$, g_i is the composition of $\boldsymbol{\Omega_{12}}$ and ψ, that is $g_i = \boldsymbol{\Omega_{12}(i)^T}\boldsymbol{\psi(\cdot)}$. They're calculated with kernel evaluation in original input space.

Notice the knowledge from the possible causal series is used to predict the residual as a measure to count its improvement to prediction power. The proposed twin-QKLMS framework is described in Algorithm 1.

3 Experiments

As a real example, we consider the physiological bivariate data (instantaneously acquired breath rate and heart rate) of a sleep human suffering from sleep apnea. The data can be downloaded: `http://physionet.incor.usp.br/physiobank/`

Algorithm 1. Twin-QKLMS Algorithm

```
Input: ${\mathbf{x}(i) \in \mathbb{X} \subseteq \mathbb{R}^m, \mathbf{y}(i) \in \mathbb{Y} \subseteq \mathbb{R}^m, x_i \in \mathbb{R}}$.
Initialization: Choose step size $\lambda, \eta > 0$, kernel width $\sigma_f, \sigma_g > 0$, the quantization size $\varepsilon_{\mathbb{X}}$, $\varepsilon_{\mathbb{Y}} \geq 0$
    and initialize the codebook (center set) $\boldsymbol{C}_f(1) = \{\mathbf{x}(1)\}$, $\boldsymbol{C}_g(1) = \{\mathbf{y}(1)\}$ and coefficient vector:
    $\alpha(1) = [\lambda x_1], \beta(1) = [0]$.
1: while $\{\mathbf{x}(i), \mathbf{y}(i), x_i\}$ $(i > 1)$ available do
2:     Compute the output of the filter $f$ and $g$:
           $f_{i-1} = \sum_{j=1}^{size(\mathbf{C}_f(i-1))} \boldsymbol{\alpha}_j(i-1)\kappa(\boldsymbol{C}_f(i-1), \mathbf{x}(i))$
           $g_{i-1} = \sum_{j=1}^{size(\mathbf{C}_g(i-1))} \boldsymbol{\beta}_j(i-1)\kappa(\boldsymbol{C}_g(i-1), \mathbf{y}(i))$
3:     Compute the error: $\xi_i = x_i - f_{i-1}$, $\xi_i^{'} = \xi_i - g_{i-1}$
4:     Compute the distance between $\mathbf{x}(i)$ and $\boldsymbol{C}_f(i-1)$ and distance between $\mathbf{y}(i)$ and $\boldsymbol{C}_g(i-1)$:
           $dis(\mathbf{x}(i), \boldsymbol{C}_f(i-1)) = \min_{1 \leq j \leq size(\boldsymbol{C}_f(i-1))} \|\mathbf{x}(i) - \boldsymbol{C}_f(i-1)\|$
           $dis(\mathbf{y}(i), \boldsymbol{C}_g(i-1)) = \min_{1 \leq j \leq size(\boldsymbol{C}_g(i-1))} \|\mathbf{y}(i) - \boldsymbol{C}_g(i-1)\|$
5:     if $dis(\mathbf{x}(i), \boldsymbol{C}_f(i-1)) \leq \varepsilon_{\mathbb{X}}$ then
6:         Keep the codebook unchanged: $\boldsymbol{C}_f(i) = \boldsymbol{C}_f(i-1)$, and quantize $\mathbf{x}(i)$ to the closest center
           through updating the coefficient of that center $\boldsymbol{\alpha}_{j^*}(i) = \boldsymbol{\alpha}_{j^*}(i-1) + \lambda e_i$,
           where $j^* = \arg\min_{1 \leq j \leq size(\boldsymbol{C}_f(i-1))} \|\mathbf{x}(i) - \boldsymbol{C}_f(i-1)\|$
7:     else
8:         Assign a new center and corresponding new coefficient: $\boldsymbol{C}_f(i) = \{\boldsymbol{C}_f(i-1), \mathbf{x}(i)\}$ and
           $\boldsymbol{\alpha}(i) = [\boldsymbol{\alpha}(i-1), \lambda e_i]$:
9:     end if
10:    Similarly, repeat step 5-9 to update the codebook $\boldsymbol{C}_g$.
11: end while
```

database/santa-fe/ (data set B). Figure 2 clearly shows that bursts of the patient breath and cyclical fluctuations of heart rate are interdependent. Both time series have been normalized to be zero mean and unit variance in preprocessing.

We set the quantization size $\varepsilon_{\mathbb{X}}$=1 and $\varepsilon_{\mathbb{Y}}$=0.5 to constrain the network size in a reasonable range. The other parameters are adjusted empirically to minimize the training error:

$$\epsilon_{n+1} = \frac{n}{n+1}\epsilon_n + \frac{1}{(n+1)}\xi_{n+1}^T \xi_{n+1} \tag{13}$$

To detect the causal relation from breath to heart rate we set λ=0.2, η=0.01, σ_f=1.6, σ_g=0.8; for the reverse case we take λ=0.02, η=0.02, σ_f=1.8, σ_g=1.1. We also evaluate the influence of different m on causality detection and the results are shown in Figure 3. We may observe: 1) with proper m values, the causal influence of heart rate on breath is, obviously, stronger than the reverse and this coincide with the results in [9]; 2) if m is too small(e.g. $m = 2$), the detected causal relationship is controversial to the expectation since long memory dynamic structured can not be modelled; 3) the causality measures change over time, and track the non-stationary dynamical behavior of the time series well.

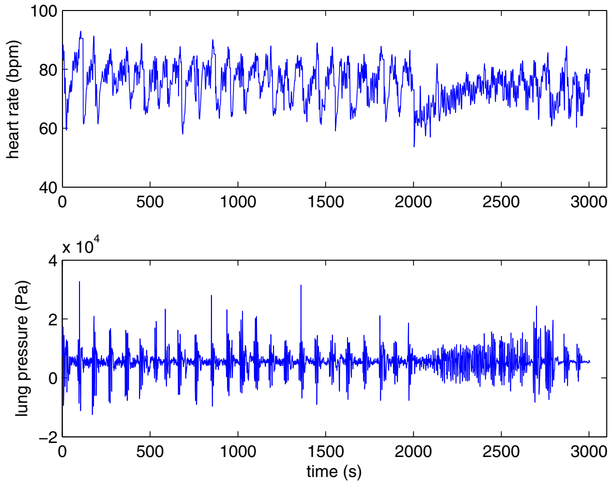
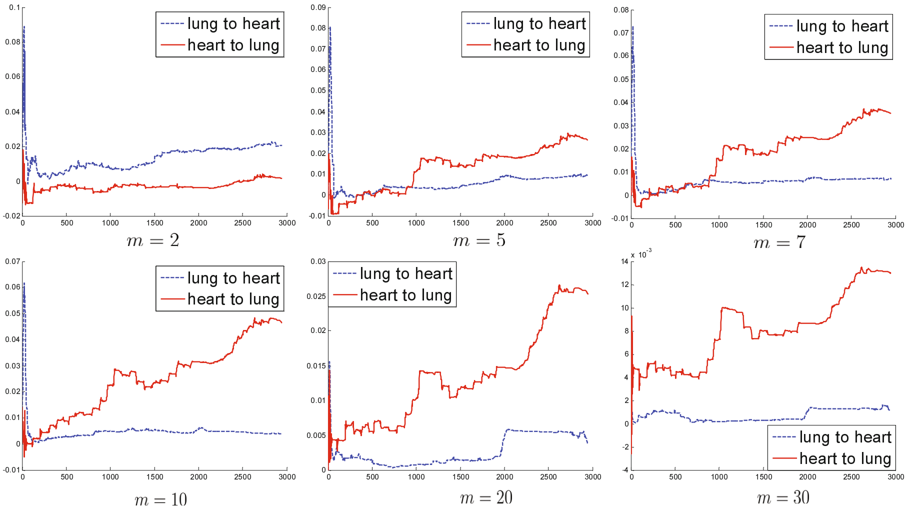


Fig. 2. Bivariate time series of the heart (upper) and breath signal (lower) of a patient suffering sleep apnea (Samples 2350-5350 of the data set are highly non-stationary and include abrupt changes in the last 1000 points). Data sampled at 2Hz.



$m = 2$ $m = 5$ $m = 7$

$m = 10$ $m = 20$ $m = 30$

Fig. 3. Causality detection results by twin-QKLMS with different m values

4 Conclusions

We develop a computationally simple online algorithm with kernel method, called twin-QKLMS, to capture instantaneous causal relationships, which can be applied especially to nonlinear and non-stationary signals. There are several key problems that need to be addressed in the future: a) the current QKLMS algorithm does not discard old centers and hence the network size is growing especially in non-stationary situation, and this leads to increase computational burden; b) how to optimize the free parameters in twin-QKLMS, since they have significant influence on the causality detection result; c) more experiments need to be done to explore its applicability such as the causality analysis in neuroscience.

References

1. Granger, C.W.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica: Journal of the Econometric Society 147, 424–438 (1969)
2. Rodriguez, E., George, N., Lachaux, J.P.: Perception's shadow: long-distance synchronization of human brain activity. Nature 397, 430–433 (1999)
3. Cadotte, A.J., DeMarse, T.B., He, P.: Causal measures of structure and plasticity in simulated and living neural networks. PloS one 3, e3355 (2008)
4. Keil, A., Sabatinelli, D., Ding, M.Z.: Re-entrant projections modulate visual cortex in affective perception: Evidence from Granger causality analysis. Human Brain Mapping 30, 532–540 (2009)
5. Akselrod, S., Gordon, D., Madwed, J.B.: Hemodynamic regulation: investigation by spectral analysis. American Journal of Physiology-Heart and Circulatory Physiology 249, H867–H875 (1985)
6. Wiener, N.: Modern mathematics for engineers. McGraw-Hill, New York (1956)
7. Seth, A.K.: A MATLAB toolbox for Granger causal connectivity analysis. Journal of Neuroscience Methods 186, 262–273 (2010)
8. Seth, A.K.: Measuring autonomy and emergence via Granger causality. Artificial Life 16, 179–196 (2010)
9. Ancona, N., Marinazzo, D., Stramaglia, S.: Radial basis function approach to nonlinear Granger causality of time series. Physical Review E 70, 056221 (2004)
10. Marinazzo, D., Liao, W., Chen, H.F.: Nonlinear connectivity by Granger causality. Neuroimage 58, 330–338 (2011)
11. Chen, B.D., Zhao, S.L., Zhu, P.P.: Quantized kernel least mean square algorithm. IEEE Transactions on Neural Networks and Learning Systems 23, 22–31 (2012)
12. Aronszajn, N.: Theory of reproducting kernels. Transactions of the American Mathematical Society, 337–404 (1950)
13. Burges, C.J.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2, 121–167 (1998)
14. Steinwart, I.: On the influence of the kernel on the consistency of support vector machines. The Journal of Machine Learning Research 2, 67–93 (2001)
15. Liu, W.F., Pokharel, P., Príncipe, J.C.: The kernel least mean square algorithm. IEEE Transactions on Signal Processing 56, 543–554 (2008)