Measures of Causality in Complex Datasets with application to financial data

Anna Zaremba 1,* and Tomaso Aste 1

1 Department of Computer Science, University College London, Gower Street, London, WC1E 6BT, UK.

* Author to whom correspondence should be addressed; A.Zaremba(at)cs.ucl.ac.uk

Abstract: This article investigates causality structure of financial time series. We concentrate on three main approaches to measuring causality: linear Granger causality, kernel generalisations of Granger causality (based on ridge regression and Hilbert-Schmidt norm of the cross-covariance operator) and transfer entropy, examining each method and comparing their theoretical properties, with special attention given to the ability to capture nonlinear causality. We also analyse the theoretical benefits of applying non symmetrical measures rather than symmetrical measures of dependence. We applied the measures to a range of simulated and real data. The simulated data sets have been generated with linear and several types of nonlinear dependence, using bivariate as well as multivariate setting. Application to real-world financial data highlights the practical difficulties as well as the potential of the methods. We use two sets of real data: (1) US inflation and 1 month Libor, (2) S&P data and exchange rates for the following currencies: AUDJPY, CADJPY, NZDJPY, AUDCHF, CADCHF, NZDCHF. Overall, we reached the conclusion that no single method can be recognised as the best in all circumstances and each of the methods has its domain of best applicability. We also describe the areas for improvement and future research.

Keywords: causality; Granger causality; Geweke’s measure of causality, transfer entropy; ridge regression; cross-covariance operator
1. Introduction

Understanding the dependence between time series is crucial for virtually all complex systems studies and the ability to describe causality structure in financial data can be very beneficial to financial institutions.

This paper concentrates on four measures of what could be referred to as “statistical causality”. There is an important distinction between the “intervention based causality” as introduced by Pearl ([1]) and “statistical causality” as developed by Granger ([2]). The first concept combines statistical and non statistical data and allows to answer questions like “if we give a drug to a patient – i.e. intervene – will the chances of their survival increase?”. The statistical causality does not answer such questions because it doesn’t operate on the concept of intervention and only allows tools of data analysis. Therefore the causality in statistical sense is a type of dependence where we infer direction as a result of the knowledge of temporal structure and the notion that the cause has to precede the effect. It can be useful for financial data, because it is commonly modelled as a single realisation of a stochastic process - a case where we cannot talk about intervention in the sense that is used by Pearl.

We will be saying that $X$ causes $Y$ in the sense of statistical (Granger) causality if the future of $Y$ can be better explained with the past of $Y$ and $X$ rather than the past of $Y$ only. We will expand and further formalise this concept using different models.

To quote Pearl “Behind every causal claim there must lie some causal assumption that is not discernable from the joint distribution and, hence, not testable in observational studies” ([1], p 40). Pearl emphasises the need to clearly distinguish between the statistical and causal terminology, and while we don’t follow his nomenclature we agree that it’s important to remember that statistical causality is not capable of discovering the “true cause”. Statistical causality can be thought of as a type of dependence and some of the methods used for describing statistical causality derive from methods used for testing for independence.

The choice of the most useful method of describing causality has to be based on the characteristics of the data and the more we know about the data the better choice we can make. In the case of financial data the biggest problems are lack of stationarity and noise. If we also consider likely nonlinearity in the dependence, model selection becomes an important factor that needs to be better understood.

The goal of the paper is to provide a broad analysis of several of the existing methods of quantifying causality. The paper is organised as follows: In section 2 we provide all the background information on the methods as well as the literature review, in section 3 we describe practical aspects, details of implementation as well as methodology of testing and the results of testing on synthetic data; financial and other application are described in the section 4; in section 5 we provide a discussion of the methods, applications and perspectives; Section 6 contains brief summary; In appendices A - E, we provide all the supplementary material that might be useful for readers not familiar with some of the concepts.

2. Methodology - literature review

2.1. Definitions of causality, methods
The first mention of causality as a property that can be estimated has appeared in 1956 in a paper by Wiener [3]: “For two simultaneously measured signals, if we can predict the first signal better by using the past information from the second one than by using the information without it, then we call the second signal causal to the first one.”

The first practical implementation of the concept was introduced by Clive Granger, the 2003 Nobel prize winner in economics, in [4] and [5]. The context in which Granger defined causality was that of linear autoregressive models of stochastic processes. Granger has described the main properties that a cause should have: it should occur before the effect and it should contain unique information about the effect that is not contained in other variables. In his works Granger included in-depth discussion of what causality means and how the statistical concept he introduced differed from deterministic causation.

2.1.1. Granger causality

In most general sense, we can say that a first signal causes a second signal if the second signal can be better predicted when the first signal is considered. It can be called Granger causality if the notion of time is introduced, so that the first signal precedes the second one. In the case when those two signals are simultaneous we will use the term instantaneous coupling.

Expanding on the original idea of Granger, the two studies published by Geweke in 1982 [6] and in 1984 [7] included the idea of feedback and instantaneous causality (instantaneous coupling). Geweke defined indices that measure the causality and instantaneous coupling with and without side information. While the indices introduced by Geweke are but one of a few alternatives that are used for quantifying Granger causality, these papers and the measures introduced therein are crucial for our treatment of causality. Geweke defined the measure of linear feedback, in place of the strength of causality used by Granger, which is one of the alternative Granger causality measures that is prevalent in the literature.

We will use notation and definitions that derive from [8], but generalise it. Let \{X_t\}, \{Y_t\}, \{Z_t\} be three stochastic processes. For any of the time series, using subscript \(t\), as in \(X_t\), will be understood as a random variable associated with the time \(t\), while using superscript \(t\), example: \(X^t\), will be understood as the collection of random variables up to time \(t\). We will use \(x_t\) and \(x^t\) as realisations of those random variables.

**Definition 1** (Granger causality) \(Y\) does not Granger cause \(X\), relative to the side information \(Z\), if for all \(t \in \mathbb{Z}\)

\[
P\left(X_t \mid X^{t-1}, Y^{t-k}, Z^{t-1}\right) = P\left(X_t \mid X^{t-1}, Z^{t-1}\right),
\]

where \(k\) is any natural number and \(P(\cdot \mid \cdot)\) stands for conditional probability distribution. If \(k = 0\) we say that \(Y\) does not instantaneously cause \(X\) (instantaneous coupling):

\[
P\left(X_t \mid X^{t-1}, Y^t, Z^{t-1}\right) = P\left(X_t \mid X^{t-1}, Z^{t-1}\right).
\]

---

1A comment on the nomenclature used: the term “Granger causality” has a range of different meanings in the literature: from a statistical test to a synonym of what we call “statistical causality”. In this paper we use the term “Granger causality” primarily as a name for a concept of a specific type of dependence, but we refer to specific measures whenever we quantify that dependence.
In the bivariate case the side information \(Z_{t-1}\) will simply be omitted. The proposed way of defining instantaneous coupling is practical to implement, but is only one of several alternative definitions, none of which is a priori superior. Amblard ([9]) recommends including \(Z_t\) rather than \(Z_{t-1}\) to ensure that the measure precludes confusing\(^2\) the instantaneous coupling of \(X\) and \(Y\) with that of \(X\) and \(Z\). Definition 1 is very general and doesn’t impose how the equality of the distributions should be assessed. The original Granger’s formulation of causality is in terms of the variance of residuals for the least-squares predictor ([5]). There are many ways of testing that; here we will, to a large degree, follow the approach from [8].

Let us here start by introducing the measures of (Granger) causality that were originally proposed by Geweke in [6] and [7]. Let \(\{X_t\}, \{Y_t\}\) be two univariate stochastic processes and \(\{Z_t\}\) be a multivariate process (the setting can be generalised to include multivariate \(\{X_t\}, \{Y_t\}\)). We assume a vector autoregressive representation, hence we assume that \(\{X_t\}\) can be modelled in the following general way:

\[
X_t = L_X(X_{t-1}) + L_{YX}(Y_{t-1}) + L_{ZX}(Z_{t-1}) + \varepsilon_{X,t} \tag{3}
\]

where \(L\) denotes a linear function and, in this particular case, it will be a linear combination. In the equation (3) we allow some of the functions \(L\) to be equal to zero everywhere. For example, if we fit \(\{X_t\}\) with the model (3) without any restrictions, and with the same model but adding the restrictions that \(L_{YX} = 0\), then we can quantify usefulness of including \(\{Y_t\}\) in explaining \(\{X_t\}\):

\[
X_t = \tilde{L}_X(X_{t-1}) + \tilde{L}_{ZX}(Z_{t-1}) + \tilde{\varepsilon}_{X,t} \tag{4}
\]

Referring to the two model representations from equations 3 and 4, Geweke’s measure of causality is formulated as follows:

\[
F_{Y \to X \parallel Z} = \log \frac{\text{var}(X_t \mid X_{t-1}, Z_{t-1})}{\text{var}(X_t \mid X_{t-1}, Y_{t-1}, Z_{t-1})} = \log \frac{\text{var}(\tilde{\varepsilon}_{X,t})}{\text{var}(\varepsilon_{X,t})}. \tag{5}
\]

Analogously, Geweke’s measure of instantaneous causality (instantaneous coupling) has been defined as:

\[
F_{Y \cdot X \parallel Z} = \log \frac{\text{var}(X_t \mid X_{t-1}, Z_{t-1})}{\text{var}(X_t \mid X_{t-1}, Y^t, Z_{t-1})} = \log \frac{\text{var}(\tilde{\varepsilon}_{X,t})}{\text{var}(\hat{\varepsilon}_{X,t})}, \tag{6}
\]

where \(\hat{\varepsilon}_{X,t}\) corresponds to yet another possible model, where both past and present of \(Y\) is considered:

\[
X_t = \hat{L}_X(X_{t-1}) + \hat{L}_{YX}(Y_t) + \hat{L}_{ZX}(Z_{t-1}) + \hat{\varepsilon}_{X,t}.
\]

In later section we will present generalisation of Geweke’s measure using kernel methods.

2.1.2. Kernels

Building on the Geweke’s linear method of quantifying causality we will introduce a nonlinear measure that uses the ‘kernel trick’, a method from machine learning to generalize liner models.

\(^2\)Yet such confusion might not be disadvantageous. If we interpret instantaneous coupling as sharing common information, we might be interested to learn that \(X\) shares contemporaneous information with \(Y\) regardless of whether any of them shares contemporaneous information with \(Z\). Without the directionality coming from the temporal structure, we cannot distinguish direct from indirect effect – which is one of the rationales for including side information when measuring causality.
There is a rich body of literature on causal inference from machine learning perspective. Initially the interest concentrated on testing for independence ([10], [11], [12], [13], [14]) but later it was recognised that independence and non-causality are related and the methods for testing one could be applied for testing the other ([15], [16]).

In particular, kernel methods have attracted much attention. In the last several years kernelisation has become a popular approach for generalising linear algorithms in many fields. The main idea underneath kernel methods is that nonlinear relationships between variables can become linear relationships between functions of the variables. This can be done by embedding (implicitly) the data into a Hilbert space enabling to search for meaningful linear relationships in that space. The main requirement of kernel methods is that the data must not be represented individually but only in terms of pairwise comparisons between the data points. Being a function of two variables, kernel function can be interpreted as a comparison function. It can also be though of as a generalization of an inner product such that the inner product is taken between functions of the variables – these functions are called ‘feature maps’.

In 2012 Pierre-Olivier Amblard and Olivier Michel published the paper [8] which, to the best of our knowledge, is the first suggesting generalisation of Granger causality using ridge regression. To some degree this late development is surprising, as ridge regression is a well established method for generalising linear regression and introducing kernels, it has a very clear interpretation, good computational properties and straightforward way of optimising parameters.

Another approach to kernelising Granger causality has been proposed by Xiaohai Sun ([17]). Sun postulated use of the square root of Hilber-Schmidt norm of the so called conditional cross-covariance operator (will be defined later as definition 3) in the feature space to measure the prediction error and permutation test to test the improvement of predictability. While none of the two kernel approaches described in this paper is based on Sun’s article, they are closely related. In particular the concept of Hilbert-Schmidt Normalised Conditional Independence Criterion (HSNCIC) is an object from a similar family as the one explored by Sun. Below, we are following the approach from [18]. Please refer to the Appendix B for supplementary information of functional analysis and Hilbert spaces.

Let us denote by \( S = (x_1, \ldots, x_n) \) a set of \( n \) observations from the process \( \{X_t\} \). We suppose that each observation \( x_i \) is an element of some set \( \mathcal{X} \). To analyse the data, and use the “kernel trick”, we create a representation of the data set \( S \) that uses pairwise comparisons \( k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \) of the points of the set \( S \) rather than the individual points. The set \( S \) is then represented by \( n \times n \) comparisons \( k_{i,j} = k(x_i, x_j) \).

**Definition 2 (Positive definite kernel)** A function \( k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \) is called a positive definite kernel iff it is symmetric, that is, \( \forall x, x' \in \mathcal{X}, k(x, x') = k(x', x) \) and (semi) positive definite, that is

\[
\forall x_1, \ldots, x_n \in \mathcal{X} \quad \forall c_1, \ldots, c_n \in \mathbb{R} \quad \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(x_i, x_j) \geq 0. \tag{7}
\]

We will use the name kernel instead of positive (semi) definite kernel henceforth.

**Theorem 1** For any kernel \( k \) on space \( \mathcal{X} \), there exist a Hilbert space \( F \) and a mapping \( \phi : \mathcal{X} \rightarrow F \) such that [18]:

\[
k(x, x') = \langle \phi(x), \phi(x') \rangle, \quad \text{for any } x, x' \in \mathcal{X}, \tag{8}
\]

where \( \langle u, v \rangle, u, v \in F \) represents an inner product in \( F \).
The above theorem leads to an alternative way of defining a kernel. It shows how we can create a kernel provided we have a feature map. Because the simplest feature map is and identity map, this theorem proves that inner product is a kernel.

The Kernel trick is a simple and general principle based on the fact that kernels can be thought of as inner products. Kernel trick can be stated as follows [18]: “Any algorithm for vectorial data that can be expressed only in terms of dot products between vectors can be performed implicitly in the feature space associated with any kernel, by replacing each dot product by a kernel evaluation.”

In the following two sections we will illustrate the use of the kernel trick in two applications: (1) the extension to nonlinear case of linear regression-Granger causality; (2) the re-formulation of concepts such as covariance and partial correlations to nonlinear case.

2.1.3. Kernelisation of Geweke’s measure of causality

Here we will show how the theory of reproducing kernel Hilbert spaces can be applied to generalise the linear measures of Granger causality as proposed by Geweke. In doing that we will use the standard theory of ridge regression.

First of all, let us go back to the model formulation of the problem that we had before (equation 3). We assumed that \( \{X_t\}, \{Y_t\} \) are two univariate stochastic processes and \( \{Z_t\} \) is a multivariate stochastic process. Let us assume we have a set of observations \( S = \{(x_1, y_1, z_1), \ldots, (x_n, y_n, z_n)\} \), \( x_t, y_t, z_t \in \mathbb{R}, z_t \in \mathbb{R}^k \). The goal is to find a function \( f \) such that for all \( t, x_t \simeq f(w_{t-1}^{t-p}) \), where \( w_i \) symbolises either \( x_i \) itself, or \( (x_i, y_i), (x_i, z_i) \), or \( (x_i, y_i, z_i) \); and \( w_{t-1}^{t-p} \) is a collection of samples \( w_i \) made from \( p \) lags prior to time \( t \), such that \( w_{t-1}^{t-p} = (w_{t-p}, w_{t-p+1}, \ldots, w_{t-1}) \). For instance, in the case where \( w \) represents all three time series: \( w_{t-1}^{t-p} = (x_{t-p}, y_{t-p}, z_{t-p}, x_{t-p+1}, y_{t-p+1}, z_{t-p+1}, \ldots, x_{t-1}, y_{t-1}, z_{t-1}) \). In general we could have that \( p \) is an infinite lag, but for any practical case it is reasonable to assume a finite lag and therefore \( w_{t-1}^{t-p} \in \mathbb{X} \), where typically \( \mathbb{X} = \mathbb{R}^d \) for \( d = p \) if \( w = x \), \( d = 2p \) if \( w = (x, y) \) and \( d = 2p + kp \) if \( w = (x, y, z) \). Using the approach of least squares regression (as in the Granger causality earlier) one would mean looking for a real valued weight vector \( \beta \) such that: \( x_t \simeq \hat{x}_t = (w_{t-1}^{t-p})^T \beta \), i.e. choosing the weight vector \( \beta \) that it minimises the squared error. The dimensionality of \( \beta \) depends on the dimensionality of \( w \), it will be a scalar in the simplest case of \( w = x \) with \( x \) being univariate.

It is well known that the drawbacks of least squares regression are poor effects with the small sample size, no solution when the problem is described by linearly dependent data and overfitting. Those problems can be addressed by adding to the cost function an additional cost penalizing excessive weights of the coefficients. This cost, called regularizer [19] or regularisation term, introduces the trade-off between the mean squared error and a norm of the weight vector. The regularised cost function is now:

\[
\beta^* = \arg\min_{\beta} \frac{1}{m} \sum_{i=p+1}^{n} ((w_{t-1}^{i-p})^T \beta - x_i)^2 + \gamma \beta^T \beta,
\]

with \( m = n - p \) for a shorter notation.

Analogously to the least squares regression weights, the solution of ridge regression (obtained in the appendix A) can be written in the form of primal weights \( \beta^* \):

\[
\beta^* = (W^T W + \gamma m I_m)^{-1} W^T x,
\]
using the matrix notation $W = \begin{pmatrix} (w^1_1)^T, (w^2_1)^T, \ldots, (w^{n-p}_{n-1})^T \end{pmatrix}^T$, or in other words a matrix with following rows: $w^1_1, w^2_{p+1}, \ldots, w^{n-p}_{n-1}$; $x = (x_{p+1}, x_{2t}, \ldots, x_n)^T$; $I_m$ denotes an identity matrix of size $m \times m$.

However, we want to be able to apply kernel methods, which require that the data is represented in the form of inner products rather than the individual data points. As is explained in Appendix A, the weights $\beta$ can be represented as a linear combination of the data points: $\beta = W^T \alpha$, for some $\alpha$. This second representation results in the dual solution $\alpha^*$ that can be represented with $WW^T$ and that depends on the regularizer $\gamma$:

$$\alpha^* = (WW^T + \gamma m I_m)^{-1} x, \quad (11)$$

This is where we can apply the kernel trick that will allow us to introduce kernels to the regression setting above. We introduce kernel similarity function $k$ which we will apply to elements of $W$. We will denote by $K_w$ the Gramm matrix built from evaluations of kernel functions on each row of $W$:

$$(K_w)_{i,j} = k(w^i_{i-1}, w^j_{j-1}), \quad \text{for } i,j = p + 1 \cdots n. \quad (12)$$

The kernel function $k$ has associated linear operator $k_w = k(\cdot, w)$. Using the representer theorem (Appendix B) allows us to represent the result of our minimisation (9) as a linear combination of kernel operators [8]. The optimal prediction can now be written in terms of the dual weights in the following way:

$$\hat{x}_t = k_w(w^t_{t-1})^T (K_w + \gamma m I_m)^{-1} x. \quad (13)$$

The mean square prediction error can be calculated by averaging over the whole set of realisations:

$$\text{var}(X_t | W^{t-1}) = \frac{1}{m} \sum_{j=1}^{l} (x_j - \hat{x}_j)^2 = \frac{1}{m} (K_w \alpha^* - x)^T (K_w \alpha^* - x), \quad (14)$$

where $\hat{x}_j$ denotes a fitted value of $x_j$.

Analogously to the Geweke’s indices from the equations (5) we can define kernelized Geweke’s indices for causality and instantaneous coupling using the above framework:

$$G_{Y \rightarrow X \| Z} = \log \frac{\text{var}(X_t | X^{t-1}, Z^{t-1})}{\text{var}(X_t | X^{t-1}, Y^{t-1}, Z^{t-1})}$$

$$G_{Y, X \| Z} = \log \frac{\text{var}(X_t | X^{t-1}, Y^{t-1}, Z^{t-1})}{\text{var}(X_t | X^{t-1}, Z^{t-1})} \quad (15)$$

### 2.1.4. Hilbert Schmidt Normalized Conditional Independence Criterion

Covariance can be used to analyse second order dependence and in the special case of variables with Gaussian distributions zero covariance is equivalent to independence. In 1959 Renyi [20] stated that to
assess independence between random variables $X$ and $Y$ one can use **maximum correlation** $S$ defined as follows:

$$S(X, Y) = \sup_{f,g} \left( \text{corr}(f(X), g(Y)) \right),$$

(16)

where $f$ and $g$ are any Borel-measurable functions for which $f(X)$ and $g(Y)$ have finite and positive variance. Maximum correlation has all of the properties than Renyi postulated for an appropriate measure of dependence, most importantly that it equals 0 if and only if the variables $X$ and $Y$ are independent. However, the concept of maximum correlation is not practical, there might not even exist such functions $f_0$ and $g_0$ for which the maximum can be attained [20]. Nevertheless, this concept has been used as a foundation of some kernel based methods for dependence, such as Kernel Constrained Covariance [21].

This section requires some background from functional analysis and machine learning. The definitions of Hilbert-Schmidt norm and operator, tensor product and mean element are given in the appendix B and follow [11] and [13].

The concept of cross-covariance operator is analogous to covariance matrix, but is defined for feature maps:

**Definition 3** (Cross-covariance operator) A linear operator $\Sigma_{XY} : \mathcal{H}_Y \rightarrow \mathcal{H}_X$ associated with the joint measure $P_{XY}$ is defined as

$$\Sigma_{XY} := E_{XY}[(\phi(X) - \mu_X) \otimes (\phi(Y) - \mu_Y)] = E_{XY}[\phi(X) \otimes \phi(Y)] - \mu_X \otimes \mu_Y,$$

(17)

where we use symbol $\otimes$ for tensor product and $\mu$ for mean embedding (definitions in Appendix B). Cross-covariance operator applied to two elements of $\mathcal{H}_X$ and $\mathcal{H}_Y$ gives the covariance:

$$\langle f, \Sigma_{XY} g \rangle_{\mathcal{H}_X} = \text{Cov}(f(X), g(Y)).$$

(18)

The notation and assumptions follow [11] and [17]: $\mathcal{H}_X$ denotes the Reproducing Kernel Hilbert Space (RKHS) induced by a strictly positive kernel $k_X : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, analogously for $\mathcal{H}_Y$ and $k_Y$. $X$ is a random variable on $\mathcal{X}$, $Y$ is a random variable on $\mathcal{Y}$ and $(X, Y)$ is a random vector on $\mathcal{X} \times \mathcal{Y}$. We assume $\mathcal{X}$ and $\mathcal{Y}$ are topological spaces and the measurability is defined with respect to the adequate $\sigma-$fields. The marginal distributions are denoted by $P_X$, $P_Y$ and the joint distribution of $(X, Y)$ by $P_{XY}$. The expectations $E_X$, $E_Y$ and $E_{XY}$ denote the expectations over $P_X$, $P_Y$ and $P_{XY}$, respectively. To ensure $\mathcal{H}_X$, $\mathcal{H}_Y$ are included in, respectively $L^2(P_X)$ and $L^2(P_Y)$, we consider only such random vectors $(X, Y)$ that the expectations $E_X[k_X(X, X)]$ and $E_Y[k_Y(Y, Y)]$ are finite.

Just as cross-covariance operator is related to the covariance, we can define an operator that is related to partial correlation:

**Definition 4** (Normalised conditional cross-covariance operator [13]) Using the cross-covariance operators we can define the normalised conditional cross-covariance operator in the following way:

$$V_{XY|Z} = \Sigma_{XX}^{-1/2} (\Sigma_{XY} - \Sigma_{XZ} \Sigma_{ZZ}^{-1/2} \Sigma_{ZY}) \Sigma_{YY}^{-1/2}.$$

(19)
Gretton et al. [11] state that for rich enough RKHS zero norm of the cross-covariance operator does imply independence, which can be written as

\[ X \perp \perp Y \iff \Sigma_{XY} = 0, \tag{20} \]

where 0 denotes a null operator. This equivalence is the premise from which follows usage of the Hilbert-Schmidt independence criterion (HSIC) as a measure of independence. Please refer to Appendix C for the information about HSIC.

It has been shown in [13] that there is a relationship similar to (20) between the normalised conditional cross-covariance operator and conditional independence, which can be written as:

\[ X \perp \perp Y \mid Z \iff V_{(XZ)(YZ)Z} = 0, \tag{21} \]

where by \((YZ)\) and \((XZ)\) we denote extended variables. Therefore the Hilbert-Schmidt norm of the conditional cross-covariance operator has been suggested as a measure of conditional independence. Using the normalised version of the operator has the advantage that it is less influenced by the marginals than non-normalised operator while retaining all the information about dependence. This is by analogy to the difference between correlation and covariance.

**Definition 5 (Hilbert Schmidt Normalised Conditional Independence Criterion – HSNCIC)** We define the HSNCIC as the squared Hilbert Schmidt norm of the normalised conditional cross-covariance operator:

\[ HSNCIC := \|V_{(XZ)(YZ)Z}\|_{HS}^2, \tag{22} \]

where \(\| \cdot \|_{HS} \) denotes Hilbert-Schmidt norm of an operator, defined in the appendix B.

For the sample \( S = \{(x_1, y_1, z_1), \ldots, (x_n, y_n, z_n)\} \) HSNCIC has an estimator that is both straightforward and has good convergence behaviour ([13], [23]). As shown in the appendix D, it can be obtained by defining empirical estimates of all of the components in following steps: first define mean elements \( \hat{m}_{X}^{(n)} \) and \( \hat{m}_{Y}^{(n)} \) and use them to define empirical cross-covariance operator \( \hat{\Sigma}_{XY}^{(n)} \). Subsequently using \( \hat{\Sigma}_{XY}^{(n)} \), together with \( \hat{\Sigma}_{XX}^{(n)} \) and \( \hat{\Sigma}_{YY}^{(n)} \) obtained in the same way, define \( \hat{V}_{XY}^{(n)} \) for the empirical normalised cross-covariance operator. Note that \( \hat{V}_{XY} \) requires inverting \( \Sigma_{YY} \) and \( \Sigma_{XX} \), hence to ensure invertibility a regularizer \( n\lambda I_n \) is added. The next step is to construct the estimator \( \hat{V}_{XY}^{(n)} \mid Z \) from \( \hat{V}_{XY}^{(n)} \), \( \hat{V}_{XZ}^{(n)} \) and \( \hat{V}_{ZY}^{(n)} \). Finally, construct the estimator of the Hilbert-Schmidt norm of \( \hat{V}_{XY}^{(n)} \mid Z \) as follows:

\[ HSNCIC_n := Tr[R(XZ)R(YZ) - 2R(XZ)R(YZ)R_Z + R(XZ)R_ZR(YZ)R_Z], \tag{23} \]

where \( Tr \) denotes a trace of a matrix, and \( R_U = K_U(K_U + n\lambda I)^{-1} \) and \( K_U(i, j) = k(u_i, u_j) \) is a Gram matrix. This estimator depends on the regularisation parameter \( \lambda \) which, in turn, depends on the sample size. Regularisation becomes necessary when inverting finite rank operators.

3By “rich enough” we mean universal, i.e. dense in the sense of continuous functions on \( X \) with the supremum norm [22].
2.1.5. Transfer entropy

Let us now introduce an alternative nonlinear information-theoretic measure of causality which is widely used and provides us with an independent comparison for the previous methods.

In 2000 Schreiber suggested measuring causality as an information transfer, in the sense of information theory. He called this measure “transfer entropy” [24]. It has become popular among physicists and biologists and there is a large body of literature on application of transfer entropy to neuroscience. We refer to [25] for a description of one of the best developed toolboxes for estimating transfer entropy. A comparison of transfer entropy and other methods of measuring causality in bivariate time series – including extended Granger causality, nonlinear Granger causality, predictability improvement and two similarity indices – has been presented by Max Lungarella et al. in [26]. A particularly exhaustive review of the relation between Granger causality and directed information has been written by Amblard et al. [9], while for a treatment of the topic from network theory perspective please refer to Amblard and Michel [27].

Transfer entropy has been designed to measure departure from generalized Markov property stating that
\[ P(X_t \mid X^{t-1}, Y^{t-1}) = P(X_t \mid X^{t-1}) \]
From the definition of Granger causality (1) for the bivariate case, i.e. with omitted side information \( \{Z_t\} \), we can see that Granger non-causality should imply zero transfer entropy (proved by Barnett et al. [28] for linear dependence of Gaussian variables and for Geweke’s formulation of Granger causality).

Transfer entropy is related to and can be decomposed in terms of Shannon entropy, as well as in terms of Shannon mutual information:

**Definition 6** (Mutual information) Assume that \( U, V \) are discrete random variables with probability distributions \( p(u_i), p(v_i) \) and joint distribution \( p(u_i, v_i) \). Then the Mutual Information \( I(U, V) \) is defined as:
\[
I(U, V) = \sum_{i,j} p(u_i, v_j) \log \frac{p(u_i, v_j)}{p(u_i)p(v_j)} = H(U) - H(U \mid V),
\]
with \( H(U) = -\sum_i p(u_i) \log p(u_i) \) the Shannon entropy and \( H(U \mid V) = \sum_{i,j} p(u_i, v_j) \log \frac{p(v_j)}{p(u_i, v_j)} \) the Shannon conditional entropy.

For independent random variables mutual information is zero. Therefore the interpretation of mutual information is that it can quantify the lack of independence between random variables and what is particularly appealing is that it does so in a nonlinear way. But being a symmetrical measure, mutual information cannot provide any information about direction of dependence. A natural extension of mutual information to include directional information is transfer entropy. According to Schreiber, the family of Shannon entropy measures are properties of static probability distribution, while transfer entropy is a generalisation to more than one system and is defined in terms of transition probabilities [24].

We assume that \( X, Y \) are random variables. As previously, \( X_t \) stands for a value at point \( t \) and \( X^t \) for a collection of values up to point \( t \).

**Definition 7** (Transfer entropy) The transfer entropy \( T_{Y \rightarrow X} \) is defined as:
\[
T_{Y \rightarrow X} = H(X_t \mid X^{t-1}) - H(X_t \mid X^{t-1}, Y^{t-1})
\]
Transfer entropy can be obtained for a multivariate system, for example [28] defines conditional transfer entropy $T_{Y \rightarrow X \mid Z} = H(X_t \mid X^t, Z^t) - H(X_t \mid X^t, Y^t, Z^t)$. We will be calculating transfer entropy only in the case of two variables. This is because the calculations already involve estimation of joint distribution of three variables $(X_t, X^t, Y^t)$ and estimating joint distribution of more variables would be very impractical for time series of length that we expect to work with.

3. Testing

3.1. Permutation tests

It has to be emphasized that in the general case the causality measures should not be used as absolute values but rather serve the purpose of comparison. While we observe that on average increasing the strength of coupling increases the value of causality, there is a large deviation in results unless the data has been generated with linear dependence and small noise. Consequently, we need a way of assessing the significance of the measure as a way of assessing significance of the causal relationship itself. For that we will be using permutation tests, following the approach in [8], [17] and [23].

By permutation test we mean a type of statistical significance test in which we use random permutations to obtain the distribution of the test statistic under the null hypothesis. For example given the aim of distinguishing between the null hypothesis $H_0$ of no causality and the hypothesis $H_1$ of causality for kernelised Geweke’s measure with side information:

$$H_0 : G_{Y \rightarrow X \mid Z} = 0$$
$$H_1 : G_{Y \rightarrow X \mid Z} > 0$$

(26)

We would like to compare the value of our causality measure on the analysed data and on appropriate “random” data and conclude that the former is significantly higher. However, generating surrogate data for such a test is not trivial and requires assumptions that affect the test results. We expect that destroying the time ordering should also destroy any potential causal effect, since statistical causality relies on the notion of time. Therefore we create the distribution of $H_0$ by reshuffling $y$, while keeping the order of $x$ and $z$ intact. More precisely, let $\pi_1, \ldots, \pi_{n_r}$ be a set of random permutations. Then instead of $y_t$ we consider $y_{\pi_j(t)}$, obtaining a set of measurements $G_{Y_{\pi_j} \rightarrow X \mid Z}$ that can be used as an estimator of the null hypothesis $G_{Y \rightarrow X \mid Z}^0$. We will accept the hypothesis of causality only if, for most of the permutations, the value of the causality measure obtained on the shuffled (surrogate) data is smaller than the value of causality measure of original data. This is quantified with a p-values defined as follows:

$$p = \frac{1}{n_r} \sum_{j=1}^{n_r} 1(G_{Y_{\pi_j} \rightarrow X \mid Z} > G_{Y \rightarrow X \mid Z})$$

(27)

Depending on the number of permutations used we suggest to accept the hypothesis of causality for the level of significance equal to 0.05 or 0.01. In our experiments we were reporting either single p-values or sets of p-values for overlapping moving windows. The latter is particularly useful when analysing noisy and non-stationary data. In the cases where not much data is available we don’t believe that using
Table 1. Dependence structure of the simulated data.

(a) Correlation matrix that has been used to generate the testing data.

|   | ts1 | ts2 | ts3 | ts4 | ts5 | ts6 | ts7 | ts8 |
|---|-----|-----|-----|-----|-----|-----|-----|-----|
| ts1 | 1   | 0.7 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| ts2 | 0.7 | 1   | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| ts3 | 0.1 | 0.1 | 1   | 0.7 | 0.1 | 0.1 | 0.1 | 0.1 |
| ts4 | 0.1 | 0.1 | 0.7 | 1   | 0.1 | 0.1 | 0.1 | 0.1 |
| ts5 | 0.1 | 0.1 | 0.1 | 0.1 | 1   | 0.7 | 0.7 | 0.7 |
| ts6 | 0.1 | 0.1 | 0.1 | 0.1 | 0.7 | 1   | 0.7 | 0.7 |
| ts7 | 0.1 | 0.1 | 0.1 | 0.1 | 0.7 | 0.7 | 1   | 0.7 |
| ts8 | 0.1 | 0.1 | 0.1 | 0.1 | 0.7 | 0.7 | 0.7 | 1   |

(b) Lags at which true dependence occurs, with the interpretation that column variable causes row variable.

|   | ts1 | ts2 | ts3 | ts4 | ts5 | ts6 | ts7 | ts8 |
|---|-----|-----|-----|-----|-----|-----|-----|-----|
| ts1 | ×   | 0   |     |     |     |     |     |     |
| ts2 | 0   | ×   |     |     |     |     |     |     |
| ts3 |     |     | ×   | -1  |     |     |     |     |
| ts4 |     |     | 1   |     |     |     |     |     |
| ts5 |     |     |     |     |     |     |     |     |
| ts6 |     |     |     |     |     |     |     |     |
| ts7 |     |     |     |     |     |     |     |     |
| ts8 |     |     |     |     |     |     |     |     |

any kind of subsampling (as proposed by [17], [8] and [23]) will be beneficial as far as the power of the tests is concerned.

3.2. Testing on simulated data

Before applying the methods to real-world data it is prudent to verify that they are working for data with known and simple dependence structure. We tested the methods on a data set containing eight time series with a relatively simple causal structure at different lags and some instantaneous coupling. We used the four methods to try to capture the dependence structure as well as to figure out which lags show dependence. The data has been simulated by first generating a set of eight time series from a Gaussian distribution with correlation matrix represented in table 1a. Subsequently, some of the series have been shifted by one, two or three time steps to obtain the following “causal” relations: $x_1 \leftrightarrow x_2$ at lag 0 i.e. instantaneous coupling of the two variables, $x_3 \rightarrow x_4$ at lag 1, $x_5 \rightarrow x_6$ at lag 1, $x_5 \rightarrow x_7$ at lag 2, $x_5 \rightarrow x_8$ at lag 3, $x_6 \rightarrow x_7$ at lag 1, $x_6 \rightarrow x_8$ at lag 2, $x_7 \rightarrow x_8$ at lag 1. The network structure is shown in the figure 1, while the lags at which the causality occurs are given in the table 1b.
Figure 1. The directionality of causality between the eight simulated time series. Green lines represent causality with the arrowheads indicating direction; red line indicates instantaneous coupling.

For the purpose of the experiments described in this paper, we have used code from several sources: own Matlab code, open access Matlab toolbox for Granger causality GCCA\(^4\) [29] and open access Matlab code provided by Sohan Seth [23]\(^5\).

To calculate Geweke’s measure and kernelised Geweke’s measure we have used the same code, with linear kernel used in the former case and Gaussian kernel – in the later; the effect of regularisation on the (linear) Geweke’s measure is negligible, the results are comparable to the GCCA code with the main difference being more flexibility on the choice of lag ranges allowed by our code. Parameters for the ridge regression were either calculated with n-fold cross-validation (Appendix E) for the grid of regularizer values in the range of \([2^{-40}, \cdots, 2^{-26}]\) and kernel sizes in the range of \([2^7, \cdots, 2^{13}]\), or fixed at a pre-specified level, with no noticeable impact on the result. Transfer entropy utilises naive histogram to estimate distributions. Code for calculating HSNCIC and for performing p-value tests incorporates framework written by Seth. The framework has been altered to accommodate some new functionality; the implementation of permutation tests has also been changed from rotation to actual permutation\(^6\). In the choice of parameters for the HSNCIC we follow [23], where the size of the kernel is set up as the median inter-sample distance and regularisation is set to \(10^{-3}\). By design this measure can only analyse one lag at a time.

The goal was to uncover the causal structure without prior information, and detect the lags at which causality occurred. This has been performed by applying all three measures of causality with following sets of lags: \([1 \cdots 10], [1 \cdots 20]\), \([1 \cdots 5], [6 \cdots 10], [11 \cdots 15]\), \([1 \cdots 3], [4 \cdots 6], [7 \cdots 9]\) and finally with all four measures to single lags \([0, 1, 2, 3, 4]\). The table 2 presents part of the results: p-values for the four measures of interest for lag 1. Below we present the conclusions for each of the methods separately, with two Geweke’s measures presented together:

Geweke’s measures. Both Geweke’s measures performed similarly, which was expected as the data was simulated with linear dependencies. Causalities were correctly identified for all ranges of lags, for

\(^4\)Code can be requested from the Author’s website: http://www.sussex.ac.uk/Users/anils/aks_code.htm
\(^5\)Code available at http://www.sohanseth.com/Home/publication/causmci
\(^6\)Use of permutation more general than rotation is helpful when data is short or the analysed time windows are short.
Table 2. P-values for four measures (from top left to bottom right: Geweke’s measure, kernelised Geweke’s measure, Transfer Entropy, HSNCIC) for lag 1

| Gc  | ts1  | ts2  | ts3  | ts4  | ts5  | ts6  | ts7  | ts8  |
|-----|------|------|------|------|------|------|------|------|
| ts1 | ×    | 0.97 | 0.35 | 0.26 | 0.24 | 0.68 | 0.11 | 0.23 |
| ts2 | 0.42 | ×    | 0.88 | 0.52 | 0.37 | 0.69 | 0.14 | 0.46 |
| ts3 | 0.26 | 0.86 | ×    | 0.75 | 0.45 | 0.19 | 0.43 | 0.72 |
| ts4 | 0.14 | 0.11 | 0    | ×    | 0.24 | 0.49 | 0.41 | 0.64 |
| ts5 | 0.78 | 0.94 | 0.10 | 0.02 | ×    | 0.40 | 0.96 | 0.91 |
| ts6 | 0.96 | 0.31 | 0.62 | 0.22 | 0    | ×    | 0.04 | 1.00 |
| ts7 | 0.74 | 0.98 | 0.10 | 0.53 | 0.35 | 0    | ×    | 0.96 |
| ts8 | 0.86 | 0.70 | 0.05 | 0.63 | 0.68 | 0.87 | 0    | ×    |

| kG  | ts1  | ts2  | ts3  | ts4  | ts5  | ts6  | ts7  | ts8  |
|-----|------|------|------|------|------|------|------|------|
| ts1 | ×    | 0.92 | 0.30 | 0.25 | 0.20 | 0.74 | 0.16 | 0.16 |
| ts2 | 0.50 | ×    | 0.93 | 0.54 | 0.52 | 0.71 | 0.19 | 0.46 |
| ts3 | 0.29 | 0.88 | ×    | 0.68 | 0.48 | 0.11 | 0.38 | 0.62 |
| ts4 | 0.12 | 0.14 | 0    | ×    | 0.22 | 0.47 | 0.41 | 0.65 |
| ts5 | 0.73 | 0.93 | 0.11 | 0.04 | ×    | 0.47 | 0.99 | 0.93 |
| ts6 | 0.94 | 0.38 | 0.55 | 0.18 | 0    | ×    | 0.07 | 0.99 |
| ts7 | 0.81 | 0.92 | 0.04 | 0.55 | 0.36 | 0    | ×    | 0.95 |
| ts8 | 0.83 | 0.67 | 0.06 | 0.63 | 0.62 | 0.86 | 0    | ×    |

which the causal direction existed, including the biggest range [1-20]. For the shorter ranges \{[1 − 5], [1 − 3]\} as well as for the single lags \{0, 1, 2, 3\} the two measures reported p-values of 0 for all of the existing causal directions. This means that the measures were able to detect precisely the lags at which causal directions existed, including the lag 0, i.e. instantaneous coupling. For those ranges, with number of permutations equal 200 and at acceptance level of 0.01, the two measures were detecting only the required causalities, but would accept some spurious causalities at level of 0.05.

**Transfer entropy.** By design, this measure can only analyse one lag at a time. It is also inherently slow, and for these two reasons it will be inefficient when a wide range of lags needs to be considered. Furthermore, it cannot be used for instantaneous coupling, to detect which we applied mutual information instead. For the lags \{1, 2, 3\} transfer entropy reported 0 p-values for all the relevant causal directions. However, it accepted spurious direction 1 → 7 with p-value of 0.01. For lag \{0\} where mutual information has been applied, the instantaneous coupling \(x_1 \leftrightarrow x_2\) has been recognised correctly with p-value 0.

**HSNCIC.** Due to slowness, HSNCIC is impractical for the largest ranges of lags. More importantly, HSNCIC performs unsatisfactorily for any of the ranges of lags that contained more than a single lag. This is deeply disappointing, as the design suggests HSNCIC should be able to handle both side information and higher dimensional variables. Even for a small range [1 – 3] HSNCIC correctly recognised only the \(x_5 \rightarrow x_8\) causality. Nevertheless, it did recognise all of the causalities correctly when analysing one lag at a time, reporting p-values of 0. This, together with the results of other tests we have run, suggests that HSNCIC is unreliable for data that has more than one lag or more than two time series. HSNCIC is also not designed to detect instantaneous coupling.

From this experiment we conclude that the Geweke’s measures with linear and Gaussian kernels provide best results and are the most practical. The other two measures, transfer entropy and HSNCIC, provide good results when analysing one lag at a time. We have also reproduced most of the tests reported by [8] and [23]. Those tests included linear and nonlinear dependence, including dependence in variance, and bivariate as well as multivariate setting. Our results were similar, but with smaller number of permutations and realisations we have obtained somewhat worse results, particularly for HSNCIC.
From all of those test we have concluded that linear causality can be detected by all measures in most
cases, with the exception of HSNCIC when more lags or dimensions are present. Granger causality can
detect some nonlinear causalities, especially if they can be approximated by linear functions. Transfer
entropy will flag more spurious causalities in the case where causal effects exist for different lags. There
is no maximum dimensionality that HSNCIC can accept, in some experiments this measure performed
well for three and four dimensional problems, in others three dimensions proved to be too many.

Possibly the most important conclusion is that parameter selection turned out to be critical for
kernelised Geweke’s measure. For some tests, like the simulated 8 time series data described earlier,
size of the kernel did not play an important role, but in some cases size of the kernel was crucial in
allowing the detection of causality. However, there was no kernel size that would work for all of the
types of the data.

4. Applications

Granger causality has been introduced as an econometrical concept and for many years it was mainly
used in economic applications. After around 30 years of relatively little acknowledgement the concept
of causality started to gain significance in a number of scientific disciplines. Granger causality and its
generalisations and alternative formulations became popular particularly in the field of neuroscience,
but also climatology and physiology ([30], [27], [31], [32], [33], [34]). The methodology could be
successfully applied in those fields, particularly in neuroscience, due to the characteristics of the data
common in those fields and the fact that assumptions of Gaussian distribution and/or linear dependence
are often reasonable [35]. This is generally not the case for financial time series.

4.1. Applications to Finance and Economy

In finance and economy there are many tools devoted to modelling dependence, mostly relevant only
for symmetrical dependence, such as correlation/covariance, cointegration, copula, to smaller degree
mutual information ([36], [37], [38], [39]). However, in various situations where we would like to
reduce the dimensionality of a problem (eg. choose subset of instruments to invest in, choose a subset of
variable for a factor model, etc.) knowledge of causality structure can help in choosing the most relevant
dimensions. Also, forecasting using the causal time series (or Bayesian priors in Bayesian models or
parents in graphical models [1], [40]) helps to forecast “future rather than the past”.

Financial data often has different characteristics than data most commonly analysed in biology,
physics etc. In finance the typical situation is that the researcher has only one long, multivariate time
series at her disposal, while in biology, even though the experiments might be expensive, most likely
there will be a number of them and usually they can be reasonably assumed independent identically
distributed (iid). The assumption of linear dependencies or Gaussian distributions, often argued to be
sensible ones in disciplines such as neuroscience, are commonly thought to be invalid for financial time
series. Furthermore, many researchers point out that stationarity usually also does not apply to this kind
of data. As causality methods in most cases do assume stationarity, the relaxation of this requirement is
clearly an important direction for future research.

In the sections below we describe the results of applying causal methods to two sets of financial data.
4.1.1. Interest rates and inflation

Interest rates and inflation have been investigated by economists for a long time. There is considerable research concerning relationship between inflation and nominal or real interest rates for the same country or region, some even utilising tools of Granger causality (for example [41]).

We are analysing related values, namely the consumer price index for the United States (US CPI) and the London Interbank Offered Rate (LIBOR) interest rate index. LIBOR is often used as a base rate (benchmark) by banks and other financial institutions and it can be thought of as an important economic indicator. It is not a monetary measure associated with any country and it does not reflect any such mandate as for example the Federal Reserve would have when setting federal interest rates. Instead, it reflects some level of assessment of risk by the banks. Therefore we ask a question: can we detect that one of these two economic indicators causes the other one in a statistical sense?

We have run our analysis for monthly data from 31 January 1986 to 31 October 2013 obtained from Thomson Reuters. Specification of implementation and parameters used for this analysis was similar as in the simulated example (section 3.2). We used kernelised Geweke’s measure with linear and Gaussian kernels. Parameters for the ridge regression were at a pre-specified level in the range of $[2^7, \cdots, 2^{13}]$ or as a median.

We investigated different time-window sizes: 25, 50, 100 and 250. The most statistically significant and interpretable results can be observed for the longer windows (250 points), where Geweke’s measure and kernelized Geweke’s measure show a clear indication of the direction US CPI $\rightarrow$ LIBOR. For shorter windows of time significant p-values were obtained considerably less often, but the results were consistent with the results for the longer time window. The dependence for the 250 days window can be seen most strongly for lag 1 (i.e. one month), less strongly for lags 2, 7, 8, 9 but there is no clear direction for the interim lags. In the figures (2) - (5) we report p-values for assessment of causality for lags 1, 2 and 7 alongside the scatter plot showing p-values and the values of Geweke’s measure. All of the charts have been scaled to show p-values in the same range $[0,1]$. We can clearly see the general trend that the higher values of causality, the lower are their corresponding p-values.
Figure 2. Kernelised Geweke’s measure of causality. The left chart shows sets of p values for the hypothesis that inflation statistically causes LIBOR (blue line) or the other way round, when model with 1 lag is considered. The right chart shows the scatter plot of p-value and value of the causality measure.

Figure 3. Linear Geweke’s measure of causality. On the left: sets of p values for the hypothesis of statistical causality in the direction US CPI → 1 month LIBOR (blue line) or the other way round (red line), when model with linear kernel and lag 1 is considered. On the right: scatter plot of p-value and value of the causality measure.
In the figure 2 we can observe that the US CPI time series lagged by one months causes 1 month LIBOR in statistical sense, when assessed with kernelised Geweke’s measure with Gaussian kernel. The p-values for hypothesis of causality in this direction allow to accept this hypothesis at significance level of 0.01 in most cases, with the p-values nearly zero most of the time. We can also observe that several of the causality measurements are as high as 0.2, which can be translated to roughly 0.18 improvement in the explanatory power of the model\(^7\). Applying linear kernel (figure 3) resulted in somewhat similar patterns of measures of causality and p-values but the two directions were less separated. Interest rates causing LIBOR still has p-values at zero most of the time, but the other direction has p-values which fall below the 0.1 level for several consecutive windows at the beginning, with the improvement in the explanatory power of the model at maximum 0.07 level; Our interpretation is that the causality is nonlinear. Please note, that kernelised Geweke’s measure with linear kernel corresponds to Geweke’s measure, hence the caption in figure 3.

**Figure 4.** Kernelised Geweke’s measure of causality. On the left: sets of p values for the hypothesis of statistical causality in the direction US CPI → 1 month LIBOR (blue line) or the other way round (red line), when model with Gaussian kernel and lag 2 is considered. On the right: scatter plot of p-value and value of the causality measure.

The results for second lag, given in figure 4, are no longer as clear as for lag 1 in figure 2 (Gaussian kernel used in both cases). The hypothesis of inflation causing interest rates still has p-values close to zero most of the time, but the p-values for the other direction are also small. This time the values of causality are lower and reach up to level below 0.08. Using linear kernel we obtain less clear results and our interpretation is that the causal direction CPI → LIBOR is stronger, but there might be some feedback as well.

\(^7\)We use Granger’s “strength of causality” for this purpose [6]. If we denote by \(\Sigma_1\) and \(\Sigma_2\) variance of errors in model 1 and 2 respectively, the strength of causality will be equal to \(1 - |\Sigma_2|/|\Sigma_1|\). The Geweke’s measure we use are given by \(ln(|\Sigma_1|/|\Sigma_2|)\), which can be transformed to give the value of strength of causality.
The figure 5 presents results of using linear kernel, which is showing a much better separation of the two directions, applied to the model with lag 7. Very similar results can be seen for models with lags 8 and 9. There is no evident reason why linear kernel performed much better than Gaussian kernel. We offer the interpretation that no nonlinear causality was strong enough and consistent enough and that this has been further obscured by using nonlinear kernel. The conclusion here is that the model selection is an important aspect of detecting causality, and needs more research.

**Figure 5.** Linear Geweke’s measure of causality. On the left: sets of p values for the hypothesis of statistical causality in the direction US CPI → 1 month LIBOR (blue line) or the other way round (red line), when model with linear kernel and lag 7 is considered. On the right: scatter plot of p-value and value of the causality measure.

In our analysis we did not obtain significant results for neither transfer entropy nor HSNCIC. The results for lag 1 for transfer entropy and HSNCIC are shown in figures respecively: 6 and 7. For lag 1 there was a significant statistical causality in the direction US CPI → 1 month LIBOR supported by both Geweke’s measures. This is barely seen for transfer entropy and HSNCIC. P-values for transfer entropy are at level that only slightly departs from a random effect and for HSNCIC they are often significant, but the two directions are not well separated. The results for higher lags were often even more difficult to interpret.
Figure 6. Transfer Entropy. On the left: sets of p values for the hypothesis of statistical causality in the direction US CPI → 1 month LIBOR (blue line) or the other way round (red line), when lag 1 is considered. On the right: scatter plot of p-value and value of the causality measure.

Figure 7. HSNCIC. On the left: sets of p values for the hypothesis of statistical causality in the direction US CPI → 1 month LIBOR (blue line) or the other way round (red line), when lag 1 is considered. On the right: scatter plot of p-value and value of the causality measure.
4.1.2. Equity versus carry trade currency pairs

We analysed six exchange rates: AUDJPY, CADJPY, NZDJPY, AUDCHF, CADCHF, NZDCHF and the index S&P and investigated any patterns of the type “leader - follower”. The expectation was that S&P should be leading. We used daily data for the period 18 July 2008 – 18 October 2013 from Thomson Reuters. We have studied the pairwise dependence between the currencies and S&P and we have also analysed the results of adding the Chicago Board Options Exchange Market Volatility Index (VIX) as a side information. In all of the cases we have used logarithmic returns.

Figure 8 presents the results of applying kernelised Geweke’s measure with Gaussian kernel. The plots show series of p-values for a moving window of length 250 data points (days), with 25 points between each window. Unlike in the previous case of interest rates and inflation, there is little actual difference between using linear and Gaussian kernel. In a few cases though employing Gaussian kernel results in better separation of the two directions, especially CADCHF $\rightarrow$ S&P and S&P $\rightarrow$ CADCHF given VIX.

Except of the CADCHF all currency pairs exhibit similar behaviour when analysed for causal effect on S&P. This behaviour consists of a small number of windows for which causal relationship is significant at p-value below 0.1 but that doesn’t persist. CADCHF is the only currency with consistently significant causal effect on S&P, which is indicated for periods starting in 2008 and 2009. As for the other direction, for AUDCHF, CADCHF and NZDCHF there are periods where S&P has a significant effect on them as measured by p-values.
**Figure 8.** Sets of p values for the hypothesis that an exchange rate causes the equity index (blue) or the other way round (red).

Figure 9 shows similar information as in figure 8, but taking into consideration VIX as a side information. The rationale is that the causal effect of S&P on the carry trade currencies is likely to be connected to the level of perceived market risk. However, the charts don’t show disappearance of causal effect after including VIX. While the patterns don’t change considerably, we observe that exchange rates have lost most of their explanatory power for S&P, with the biggest differences for CADCHF. There is little difference for the p-values for the other direction, hence the distinction between the two directions became more significant.
Figure 9. Sets of p values for the hypothesis that an exchange rate causes the equity index, given VIX as side information (blue) or the other way round (red).

We obtained all of the main “regimes”: periods when either one of the exchange rates or S&P had more explanatory power (p-values for one direction much lower than for the other), periods when both exhibited low or both exhibited high p-values. P-values close to 1 didn’t necessarily mean only lack of causality: in such cases the random permutations of the time series tested for causality at a specific lag appear to have higher explanatory power than the time series at this lag itself. There are a few possible explanations: related to the data, the measures and to the nature of the permutation test itself. We have observed on the simulated data that when no causality is present, autocorrelation introduces biases to the permutation test: higher p-values than we would expect from a randomised sample, but also higher likelihood of interpreting correlation as causality. Furthermore, both of these biases can result from assuming model with different amount of lags than the data was simulated with. Correspondingly, if
the data has been simulated with instantaneous coupling and no causality, this again can result in high p-values. Out of all four methods transfer entropy appeared to be most prone to all these biases.

Relating these observations to the exchange rates and S&P data described in this section, we believe that the high p-values result primarily from the fact that the time series are all correlated. The autocorrelation has, arguably, smaller effect because taking logarithmic returns of the time series results in lack of significant sample autocorrelations for majority of time windows.

5. Discussion and Perspectives

While questions about causal relations are asked often in science in general, the appropriate methods of quantifying causality for different contexts are not well developed. Firstly, often answers are formulated with methods not intended specifically for this purpose. There are fields of science, for example nutritional epidemiology, where causation is commonly inferred from correlation. The classical example from economy, known as “Milton Friedmans thermostat”, described confusing lack of correlation with lack of causation in evaluation of the Federal Reserve [43]. Secondly, often questions are formulated in terms of (symmetrical) dependence because it involves established methods and allows clear interpretation. This could be a case in many risk management applications where the question of what causes losses should be a central one but is not commonly addressed with causal methods [44]. The tools for quantifying causality that are currently being developed can help make that causal inference and better understand the results.

In this section we provide a critique of the methods to help understand their weaknesses and enable one to choose the most appropriate method for intended use. This will also set out possible directions of future research. The first part of this section describes the main differences between the methods, followed by a few comments on model selection and problems related to permutation testing. Indication of directions of future research conclude the section.

5.1. Theoretical differences

**Linearity versus nonlinearity.** The original Granger causality and its Geweke’s measure formulation have been developed to assess linear causality and they are very robust and efficient in doing so. For data with linear dependence using linear Granger causality is most likely to be the best choice. The measure can work well also in cases where the dependence is not linear but has a strong linear component.

As financial data does not normally exhibit stationarity, linearity and Gaussianity, arguably linear methods should not be used to analyse them. In practice, requirements on the size of the data and difficulties with model selection take precedence and mean that linear methods should still be considered.

**Direct and indirect causality.** Granger causality is not transitive, which might be unintuitive. Although transitivity would bring the causality measure closer to the common understanding of the term, it could also make it impossible to distinguish between direct and indirect cause. As a consequence it could make the measure useless for the purpose of reduction of dimensionality and repeated information. However, differentiation between direct and indirect causality is not necessarily well defined. This

---

8An example that had a considerable media coverage in many countries is the paper from 2012 by scientists from the Harvard School of Public Health about effect of consumption of red meat on mortality [42].
is because adding a conditioning variable can both introduce as well as remove dependence between variables [45]. Hence the notion of direct and indirect causality is relative to the whole information system and can change if we add new variables to the system. Using methods from graphical modelling [1] could facilitate defining the concepts of direct and indirect causality, as these two terms are well defined for causal networks.

Geweke’s and kernelised Geweke’s measures can distinguish direct and indirect cause in some cases. Following the example of Amblard [8] we suggest comparing the conditional and non-conditional causality measurements as means of distinguishing between direct and indirect cause for both linear and kernel Granger causality. Measures like HSNCIC are explicitly built in such a way that they condition on side information and therefore are geared towards picking up only the direct cause; this, however, does not work as intended as we have noticed that HSNCIC is extremely sensitive to dimensionality of the data. Transfer entropy – in the form we are using – does not consider side information at all. A new measure, called partial transfer entropy ([46], [47]) has been proposed to distinguish between direct and indirect cause.

**Spurious causality.** Partially covered in the previous point about direct and indirect cause, the problem of spurious causality is a wider one. As already indicated, causality is indicated only relative to given data and introducing more data can both add and remove (spurious) causalities. The additional problem is that data can exhibit many types of dependence. None of the methods we are discussing in this paper is capable of managing several simultaneous types of dependence, be it instantaneous coupling, linear or nonlinear causality. We refer the interested reader to relevant literature on modelling Granger causality and transfer entropy in the frequency domain or using filters [29], [48], [49].

**Numerical estimator.** It was already mentioned that Granger causality and kernel Granger causality are robust for small samples and high dimensionality. Both of those measures optimise quadratic cost, which means they can be sensitive to outliers, but kernelised Geweke’s measure can managed this with parameter selection. Granger causality for bivariate data has good statistical tests for significance, while the others have not and need permutation tests which are computationally expensive. Also, in the case of ridge regression, there is another layer of optimising parameters which is also computationally expensive. Calculating kernels is also computationally expensive (unless the data is high-dimensional), but they are robust for small samples.

HSNCIC is shown to have a good estimator which in the limit of infinite data doesn’t depend on the type of kernel. Transfer entropy, on the other hand, has estimators that are unreliable for small and medium samples. A detailed overview of possible methods of estimation of entropy can be found in [30]. Trentool [25], one of more popular open access toolboxes for calculating transfer entropy, uses nearest neighbour technique to estimate joint and marginal probabilities [25]. This technique is supposed to greatly reduce bias but it is demanding with respect to the size of the data and is slower than our simple approach. We have tested Trentool and have found that the demands on the size of the sample where too high.

**Non-stationarity.** This is one of the most important areas for future research. All of the described measures suffer to some degree from inability to deal with non-stationary data. Transfer entropy does not explicitly assume stationarity, but in practice it can still be affected if the time series is highly non-stationary.
The GCCA toolbox\(^9\) for calculating Granger causality provides some tools for detecting non-stationarity and to a limited degree also for managing it [29]. In the vector autoregressive setting of Granger causality it is possible to run parametric tests to detect nonstationarity: ADF test (Augmented Dickey Fuller) and KPSS test (Kwiatkowski, Phillips, Schmidt, Shin). For managing non-stationarity Seth suggests analysing shorter time series (windowing) and differencing, although both approaches can introduce new problems. It is also advisable to detrend and demean the data, and in the case of economic data it might also be possible to perform seasonal adjustment.

**Choice of parameters.** Each of the methods requires selection of parameters – an issue related to model selection described in the section 5.2. All of the methods need the choice of the number of lags, while kernel methods additionally require choice of kernel, kernel parameter (kernel size) and regularisation parameter.

In the case of Gaussian kernel, the effect of the kernel size on the smoothing of the data could be understood as follows ([50], [51]). The Gaussian kernel \(k(x, y) = \exp(-\|x - y\|^2/\sigma^2)\) corresponds to an infinite dimensional feature map consisting of all possible monomials of input features. For the Gaussian kernel, or more generally for any kernel that can be defined by a Taylor series expansions, if we consider the basis \(1, u, u^2, u^3, \ldots\) then the random variables \(X\) and \(Y\) can be expressed in RKHS by:

\[
\Phi(X) = k(X, u) \sim (1, c_1 X, c_2 X^2, c_3 X^3, ...)^T
\]

\[
\Phi(Y) = k(Y, u) \sim (1, c_1 Y, c_2 Y^2, c_3 Y^3, ...)^T,
\]

therefore the kernel function can be expressed as follows:

\[
k(x, u) = 1 + c_1 x u + c_2 x^2 u^2 + c_3 x^3 u^3 + ...
\]

(29)

and the cross-covariance matrix will contain information on all of the higher-order covariances:

\[
\Sigma_{XY} \sim \begin{pmatrix}
0 & 0 & 0 & 0 & 0 \\
0 & c_1 Cov[X, Y] & c_1 c_2 Cov[X, Y^2] & c_1 c_3 Cov[X, Y^3] & \ldots \\
0 & c_2 Cov[X^2, Y] & c_2^2 Cov[X^2, Y^2] & c_2 c_3 Cov[X^2, Y^3] & \ldots \\
0 & c_3 Cov[X^3, Y] & c_3 c_2 Cov[X^3, Y^2] & c_3^2 Cov[X^3, Y^3] & \ldots \\
0 & \ldots & \ldots & \ldots & \ldots
\end{pmatrix}
\]

(30)

According to Fukumizu et al. [13] the HSNCIC measure does not depend on the kernel in the limit of infinite data. However, the other parameters still need to be chosen, which is clearly a drawback. Kernelised Geweke’s measure optimises parameters explicitly with the cross-validation while HSNCIC focuses on embedding the distribution in RKHS with any characteristic kernel. Additionally, transfer entropy requires the choice of method for estimating densities, the binning size in the case of naive histogram approach.

5.2. Model selection

\(^9\)Code can be requested at: http://www.sussex.ac.uk/Users/anils/aks_code.htm
Table 3. The summary of main features of the different measures

| Measures               | Properties                                                                 |
|------------------------|-----------------------------------------------------------------------------|
|                        | **Linearity versus nonlinearity**                                           |
| Granger causality      | assumes linearity; the best method for linear data, the worst for nonlinear |
| kernelised Geweke’s   | works for both linear and nonlinear data                                   |
| transfer entropy       | works for both linear and nonlinear data                                   |
| HSNCIC                 | works for both linear and nonlinear data if low dimension                  |
|                        | **Distinguishing direct from indirect causality**                          |
| Granger causality      | to some extent by comparing measure with and without side information      |
| kernelised Geweke’s   | to some extent by comparing measure with and without side information      |
| transfer entropy       | might not be capable of                                                    |
| HSNCIC                 | to some extent, as it is designed to condition on side information         |
|                        | **Spurious causality**                                                     |
| Granger causality      | susceptible                                                                 |
| kernelised Geweke’s   | susceptible                                                                 |
| transfer entropy       | susceptible                                                                 |
| HSNCIC                 | susceptible                                                                 |
|                        | **Good numerical estimator**                                               |
| Granger causality      | yes                                                                        |
| kernelised Geweke’s   | yes                                                                        |
| transfer entropy       | no                                                                          |
| HSNCIC                 | yes                                                                        |
|                        | **Nonstationarity**                                                        |
| Granger causality      | v. sensitive; test with ADF, KPSS, use windowing, differencing, large lag  |
| kernelised Geweke’s   | somewhat sensitive; online learning is a promising approach                |
| transfer entropy       | somewhat sensitive                                                          |
| HSNCIC                 | somewhat sensitive                                                          |
|                        | **Choice of parameters**                                                   |
| Granger causality      | lag                                                                         |
| kernelised Geweke’s   | kernel, kernel size, regularisation parameter, lag; uses cross-validation   |
| transfer entropy       | lag, binning size (if histogram approach used)                              |
| HSNCIC                 | kernel, kernel size, regularisation parameter, lag                         |
For the kernel measures we observe that model selection becomes an important issue. In general, the choice of kernel influences the smoothness of the class of functions considered, while the choice of regularizer controls the trade-off between smoothness of the function and the error of the fit. Underfitting can be a consequence of too large regularizer and too large kernel size (in case of Gaussian kernel); conversely, overfitting can be a consequence of too small regularizer and too small kernel size.

One of the methods suggested to help with model selection is cross-validation [8]. Given nonstationary data it would seem reasonable to fit the parameters; we concluded, however, that cross-validation was too expensive in computational sense and did not provide the expected benefits.

We feel that more research is needed on the model selection.

5.3. Testing

Indications of spurious causality can be generated not only when applying measures of causality but also when testing those measures. The permutation test that has been described in the section 3.1 involves destruction of all types of dependence, not just causal dependence. In practice it means that existence of instantaneous coupling can result in incorrect interpretation of causal inference. Nevertheless, simplicity is the deciding factor in preference of permutation tests to other approaches.

Several authors ([23], [8], [17]) propose repeating the permutation test on subsamples to achieve acceptance rates, an approach we do not favour in practical applications. The rationale for using acceptance rates is that the loss on significance from decreasing the size of the sample will be more than made up by calculating many permutation tests for many subsamples. We believe this might be reasonable in the case where the initial sample was big and the assumption of stationarity was a reasonable one, but that was not the case for our data. We have instead decided to report p-values for an overlapping running window. This allows us to additionally assess consistency of results and does not require to choose the same significance rate for all of the windows.

5.4. Perspectives

In the discussion we have highlighted many areas which still require more research. The kernelised Geweke’s measure, transfer entropy and HSNCIC allow detecting nonlinear dependence better that the original Granger causality, but don’t improve on the other important area of its weakness: non-stationarity. Ridge regression is expected to be a convenient tool to use in online learning - methodology which could prove helpful in dealing with non-stationarity [8]. This is clearly an area worth exploring.

Crucially for applications to financial data, more should be understood about measuring causality in time series with several types of dependence. We are not aware of any study that targets this question. We believe this should be approached first by analysing synthetic models. A possible direction of research here is using filtering to prepare data before causal measures are applied. One possibility is to frequency based decomposition. A different type of filtering is decomposition into negative and positive shocks, for example Hatemi-J proposed “assymmetric causality measure” based on Granger causality [52].

The third biggest area of suggested research is building causal networks. There is a substantial body of literature about causal networks for the intervention based causality, as it is described in terms of
graphical models. Prediction based causality has been used less often to describe causal networks, but this approach is becoming more popular ([53], [41], [27]). Successfully building a complex causal network requires particular attention to side information and the distinction between direct and indirect cause. This is a very interesting area of research with various applications in finance, in particular: portfolio diversification, causality arbitrage portfolio, risk management for investments, etc.

6. Conclusions

We have compared causality measures based on methods from the fields of econometrics, machine learning and information theory. After analysing their theoretical properties and the results of the experiments we can conclude that no measure is clearly better than the others. We believe, however, that kernelised Geweke’s measure based on ridge regression is the most practical, performing relatively well for both linear and nonlinear causal structures as well as for both bivariate and multivariate systems. For the two sets of real data we were able to identify causal directions that showed some consistency between methods, time windows, and that were not contradictory with economic rationale. The two experiments allowed to identify a range of limitations that need to be addressed to allow wider application of any the methods to financial data. Also, neither of the two sets contained high frequency data and working with high frequency is likely to produce additional complications.

Separate question that we only touch upon is the relevance and practicality of using any causality measure. This is a question lying largely in the domain of philosophy o science. Ultimately, it is the interpretation of researcher and their confidence in the data that makes it possible to label a relationship as causal rather than only statistically causal. But while the measures that we analyse cannot discover a true cause or distinguish categorically between true and spurious causality, they can still be very useful in practice.

Granger causality has often been used in economic models and has gained even wider recognition after Granger received Nobel prize in 2003. There is little literature on using nonlinear generalisations of the Granger causality in finance or economy. We believe that it has large potential on one hand and still many questions to answer on the other. While we expect that some of the problems could be addressed with online learning approach and data filtering, more research on dealing with non-stationarity, noisy data and optimal parameter selection is required.
A. Solving ridge regression

The regularised cost function is (9):

\[
\beta^* = \arg\min_\beta \frac{1}{m} \sum_{i=p+1}^{n} ((w_i^{i-p})^T \beta - x_i)^2 + \gamma \beta^T \beta,
\]

(31)

Now solving 9 gives:

\[
L = \frac{1}{m} \sum_{i=p+1}^{n} ((w_i^{i-p})^T \beta - x_i)^2 + \gamma \beta^T \beta
= \frac{1}{m} (W \beta - x)^T (W \beta - x) + \gamma \beta^T \beta
= \frac{1}{m} (\beta^T W^T W \beta - 2x^T W \beta + x^T x) + \gamma \beta^T \beta
\]

(32)

\[
\frac{\partial L}{\partial \beta} = \frac{1}{m} (2W^T W \beta - W^T x) + 2\gamma \beta = 0 \iff
W^T W \beta^* + \gamma m \beta^* = W^T x \iff
\beta^* = (W^T W + \gamma m I^m)^{-1} W^T x,
\]

where \( I^m \) is an \( m \times m \) identity matrix.

The weights \( \beta^* \) are called the **primal solution** and the next step is to introduce the **dual solution** weights.

\[
W^T W \beta^* + \gamma m \beta^* = W^T x \iff \beta^* = \frac{1}{\gamma m} W^T (x - W \beta^*),
\]

(33)

so for some \( \alpha^* \in \mathbb{R}^n \) we can write that

\[
\beta^* = W^T \alpha^*.
\]

(34)

From the two sets of equation above we get that:

\[
\alpha^* = \frac{1}{\gamma m} (x - W \beta^*)
\iff \gamma m \alpha^* = x - W \beta^* = x - W W^T \alpha^* \iff
(W W^T + \gamma m I_m) \alpha^* = x.
\]

(35)

This gives the desired form for the dual weights:

\[
\alpha^* = (W W^T + \gamma m I_m)^{-1} x,
\]

(36)

which depend on the regularisation parameter \( \gamma \).

B. Definitions from functional analysis and Hilbert Spaces

The definitions and theorems below follow [54], [11] and [17]. All vector spaces will be over \( \mathbb{R} \), rather than \( \mathbb{C} \), albeit they can all be generalised for \( \mathbb{C} \) with little modification.
Definition 8 (Inner product) Let $\mathcal{F}$ be a vector space over $\mathbb{R}$. A function $\langle \cdot, \cdot \rangle : \mathcal{F} \times \mathcal{F} \rightarrow \mathbb{R}$ is said to be an inner product on $H$ if:

i) $\langle f_1 + f_2, f \rangle = \langle f_1, f \rangle + \langle f_2, f \rangle$, for all $f, f_1, f_2 \in \mathcal{F}$

ii) $\langle \alpha f_1, f_2 \rangle = \alpha \langle f_1, f_2 \rangle$ for all $f_1, f_2 \in \mathcal{F}, \alpha \in \mathbb{R}$

iii) $\langle f_1, f_2 \rangle = \langle f_2, f_1 \rangle$ for all $f_1, f_2 \in \mathcal{F}$

ii) $\langle f, f \rangle \geq 0$ and $\langle f, f \rangle = 0$ if and only if $f = 0$.

Definition 9 (Hilbert space) If $\langle \cdot, \cdot \rangle$ is an inner product on $\mathcal{F}$, the pair $(\mathcal{F}, \langle \cdot, \cdot \rangle)$ is called a Hilbert space if $\mathcal{F}$ with metric induced by the inner product is complete.

One of the fundamental concepts of functional analysis that we will utilise is that of continuous linear operator: for two vector spaces $\mathcal{F}$ and $\mathcal{G}$ over $\mathbb{R}$, a map $T : \mathcal{F} \rightarrow \mathcal{G}$ is called a (linear) operator if it satisfies $T(\alpha f) = \alpha T(f)$ and $T(f_1 + f_2) = T(f_1) + T(f_2)$ for all $\alpha \in \mathbb{R}, f_1, f_2 \in \mathcal{F}$. Throughout the rest of the paper we use standard notational convention $Tf := T(f)$.

The following three conditions can be proved to be equivalent: (1) linear operator $T$ is continuous, (2) $T$ is continuous at 0, (3) $T$ is bounded. This result together with Riesz representation theorem given later, are fundamental for the theory of Reproducing Kernel Hilbert Spaces. It should be emphasised that while the operators we will use, such as mean element and cross-covariance operator are linear, the functions they operate on will not, in general, be linear. An important special case of linear operators are the linear functionals, which are such operators $T$ that: $T : \mathcal{F} \rightarrow \mathbb{R}$.

Theorem 2 (Riesz representation theorem) In a Hilbert space $\mathcal{F}$, all continuous linear functionals are of the form $\langle \cdot, f \rangle$, for some $f \in \mathcal{F}$.

In the previous appendix A we have used the “kernel trick” without explaining why it was permissible. The explanation is given below as the Representer Theorem. The theorem will refer to a loss function $L(x, y, f(x))$ that describes the cost of the discrepancy between the prediction $f(x)$ and the observation $y$ at the point $x$. Risk $\mathcal{R}_{L,S}$ associated with the loss $L$ and data sample $S$ is defined by the average future loss of the prediction function $f$.

Theorem 3 (Representer theorem)[54] Let $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \rightarrow [0, \infty)$ be a convex loss, $S := \{x_1, y_1, ..., (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$ be a set of observations and $\mathcal{R}_{L,S}$ denote associated risk. Furthermore, let $\mathcal{F}$ be an RKHS over $\mathcal{X}$. Then for all $\lambda > 0$ there exists a unique empirical solution function which we denote by $f_{S,\lambda} \in \mathcal{F}$ satisfying the equality:

$$\lambda \|f_{S,\lambda}\|^2_{\mathcal{F}} + \mathcal{R}_{L,S}(f_{S,\lambda}) = \inf_{f \in \mathcal{F}} \lambda \|f\|^2_{\mathcal{F}} + \mathcal{R}_{L,S}(f)$$

---

10A metric space is complete if very Cauchy sequence converges in this space.

11A bounded linear operator is generally not a bounded function.

12If $\mathcal{F}$ is a normed space, the space of all continuous linear functionals $T : \mathcal{F} \rightarrow \mathbb{R}$ is called a topological dual space of $\mathcal{F}$ and denoted as $\mathcal{F}'$. 
In addition, there exist \( \alpha_1, \cdots, \alpha_n \in \mathbb{R} \) such that

\[
f_{S,\lambda}(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i), \quad \text{for } x \in \mathcal{X}. \tag{39}\]

Below we present definitions which are building blocks of the Hilbert Schmidt Normalized Conditional Independence Criterion.

**Definition 10** *(Hilbert-Schmidt norm)* Let \( \mathcal{F} \) be a Reproducing Kernel Hilbert Space (RKHS) of functions from \( \mathcal{X} \) to \( \mathbb{R} \), induced by strictly positive kernel \( k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \). Let \( \mathcal{G} \) be an RKHS of functions from \( \mathcal{Y} \) to \( \mathbb{R} \), induced by strictly positive kernel \( l : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R} \). Denote by \( C : \mathcal{G} \rightarrow \mathcal{F} \) a linear operator. The Hilbert-Schmidt norm of the operator \( C \) is defined as

\[
\|C\|_{HS}^2 := \sum_{i,j} \langle Cv_i, u_j \rangle_{\mathcal{F}}^2, \tag{40}\]

given that the sum converges, where \( u_i \) and \( u_j \) are orthonormal bases of \( \mathcal{F} \) and \( \mathcal{G} \) respectively; \( \langle v, u \rangle_{\mathcal{F}}, u, v \in \mathcal{F} \) represents an inner product in \( \mathcal{F} \).

Following [11] and [17], let \( \mathcal{H}_\mathcal{W} \) denote the RKHS induced by a strictly positive kernel \( k_\mathcal{W} : \mathcal{W} \times \mathcal{W} \rightarrow \mathbb{R} \). Let \( X \) be random variable on \( \mathcal{X} \), \( Y \) be random variable on \( \mathcal{Y} \) and \((X,Y)\) be random vector on \( \mathcal{X} \times \mathcal{Y} \). We assume \( \mathcal{X} \) and \( \mathcal{Y} \) are topological spaces and the measurability is defined with respect to the adequate \( \sigma \)-fields. The marginal distributions are denoted by \( P_X, P_Y \) and the joint distribution of \((X,Y)\) by \( P_{XY} \). The expectations \( E_X, E_Y \) and \( E_{XY} \) denote the expectations over \( P_X, P_Y \) and \( P_{XY} \), respectively. To ensure \( \mathcal{H}_X, \mathcal{H}_Y \) are included in, respectively \( L^2(P_X) \) and \( L^2(P_Y) \), we consider only such random vectors \((X,Y)\) that the expectations \( E_X[k_X(X,X)] \) and \( E_Y[k_Y(Y,Y)] \) are finite.

**Definition 11** *(Hilbert-Schmidt operator)* A linear operator \( C : \mathcal{G} \rightarrow \mathcal{F} \) is Hilbert-Schmidt if its Hilbert-Schmidt norm exists.

The set of Hilbert-Schmidt operators \( HS(\mathcal{G}, \mathcal{F}) : \mathcal{G} \rightarrow \mathcal{F} \) is a separable Hilbert space with the inner product defined:

\[
\langle C, D \rangle_{HS} := \sum_{i,j} \langle Cv_i, u_j \rangle_{\mathcal{F}} \langle Dv_i, u_j \rangle_{\mathcal{F}}, \tag{41}\]

where \( C, D \in HS(\mathcal{G}, \mathcal{F}) \).

**Definition 12** *(Tensor product)* Let \( f \in \mathcal{F} \) and \( g \in \mathcal{G} \), then the tensor product operator \( f \otimes g : \mathcal{G} \rightarrow \mathcal{F} \) is defined as follows:

\[
(f \otimes g)h := f \langle g, h \rangle_{\mathcal{G}}, \quad \text{for all } h \in \mathcal{G}. \tag{42}\]

The definition above makes use of two standard notational abbreviations. The first one concerns omitting brackets when denoting application of an operator: \( (f \otimes g)h \) instead of \( (f \otimes g)(h) \). The second one relates to multiplication by scalar and writing \( f \langle g, h \rangle_{\mathcal{G}} \) instead of \( f \cdot \langle g, h \rangle_{\mathcal{G}} \).

---

\[\text{We will require that space } F \text{ must be separable (to have a complete orthonormal system), but in practice we will use } \mathbb{R}^n \text{ and therefore that won't be an issue.}\]
The Hilbert-Schmidt norm of the tensor product can be calculated as:

$$\|f \otimes g\|_{2,HS}^2 = \langle f \otimes g, f \otimes g \rangle_{HS} = \langle f, (f \otimes g)g \rangle_{\mathcal{F}}$$

$$= \langle f, f \rangle_{\mathcal{F}} \langle g, g \rangle_{\mathcal{G}} = \|f\|_{2,\mathcal{F}}^2 \|g\|_{2,\mathcal{G}}^2.$$  (43)

When introducing the cross-covariance operator we will be using the following results for the tensor product. Given a Hilbert-Schmidt operator $L : \mathcal{G} \rightarrow \mathcal{F}$ and $f \in \mathcal{F}$ and $g \in \mathcal{G}$,

$$\langle L, f \otimes g \rangle_{HS} = \langle f, Lg \rangle_{\mathcal{F}}.$$  (44)

A special case of the equation 44 is, with the notation as earlier and $u \in \mathcal{F}$ and $v \in \mathcal{G}$,

$$\langle f \otimes g, u \otimes v \rangle_{HS} = \langle f, u \rangle_{\mathcal{F}} \langle g, v \rangle_{\mathcal{G}}.$$  (45)

**Definition 13** *(The mean element)* Given the notation as above, we define the mean element $\mu_X$ with respect to the probability measure $P_X$ as such element of the RKHS $\mathcal{H}_X$ for which

$$\langle \mu_X, f \rangle_{\mathcal{H}_X} := E_X[\langle \phi(X), f \rangle_{\mathcal{H}_X}] = E_X[\phi(X)],$$  (46)

where $\phi : \mathcal{X} \rightarrow \mathcal{H}_X$ is a feature map and $f \in \mathcal{H}_X$.

The mean elements exist as long as their respective norms are bounded, which condition will be met if the relevant kernels are bounded.

**C. Hilbert-Schmidt Independence Criterion (HSIC)**

As in the section 2.1.4, following [11] and [17], let $\mathcal{F}_X, \mathcal{F}_Y$ denote the RKHS induced by strictly positive kernels $k_X : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ and $k_Y : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$. Let $X$ be random variable on $\mathcal{X}$, $Y$ be random variable on $\mathcal{Y}$ and $(X, Y)$ be random vector on $\mathcal{X} \times \mathcal{Y}$. The marginal distributions are denoted by $P_X, P_Y$ and the joint distribution of $(X, Y)$ by $P_{XY}$.

**Definition 14** *(Hilbert-Schmidt Independence Criterion – HSIC)* With the notation for $\mathcal{F}_X, \mathcal{F}_Y, P_X, P_Y$ as introduced earlier, we define the Hilbert Schmidt independence criterion as the squared Hilbert Schmidt norm of the cross-covariance operator $\Sigma_{XY}$:

$$HSIC(P_{XY}, \mathcal{F}_X, \mathcal{F}_Y) := \|\Sigma_{XY}\|_{HS}^2.$$  (47)

We cite without proof the following lemma from [11]:

**Lemma 1** *(HSIC in kernel notation)*

$$HSIC(P_{XY}, \mathcal{F}_X, \mathcal{F}_Y) := E_{X,X',Y,Y'}[k_X(X, X')k_Y(Y, Y')] + E_{X,X'}[k_X(X, X')]E_{Y,Y'}[k_Y(Y, Y')]$$

$$-2E_{X,Y}[E_{X'}[k_X(X, X')]E_{Y'}[k_Y(Y, Y')]].$$  (48)

where $X, X'$ and $Y, Y'$ are independent copies of the same random variable.
D. Estimator of HSNCIC

Empirical mean elements:

\[
\hat{m}_X^{(n)} = \frac{1}{n} \sum_{i=1}^{n} k_X(\cdot, X_i),
\]
\[
\hat{m}_Y^{(n)} = \frac{1}{n} \sum_{i=1}^{n} k_Y(\cdot, Y_i)
\]

(49)

Empirical cross-covariance operator:

\[
\hat{\Sigma}_{XY}^{(n)} = \frac{1}{n} \sum_{i=1}^{n} (k_Y(\cdot, Y_i) - \hat{m}_Y^{(n)}) \langle k_X(\cdot, X_i) - \hat{m}_X^{(n)}, \cdot \rangle_{\mathcal{H}_X}
\]
\[
= \frac{1}{n} \sum_{i=1}^{n} \{ k_Y(\cdot, Y_i) - \hat{m}_Y^{(n)} \} \otimes \{ k_X(\cdot, X_i) - \hat{m}_X^{(n)} \}
\]

(50)

Empirical normalised cross-covariance operator

\[
\hat{V}_{XY}^{(n)} = (\hat{\Sigma}_{XX}^{(n)} + n\lambda I_n)^{-1/2} \hat{\Sigma}_{XY}^{(n)} (\hat{\Sigma}_{YY}^{(n)} + n\lambda I_n)^{-1/2}
\]

(51)

where $n\lambda I_n$ is added to ensure invertibility.

Empirical normalised conditional cross-covariance operator

\[
\hat{V}_{XY|Z}^{(n)} = \hat{V}_{XY}^{(n)} - \hat{V}_{XZ}^{(n)} \hat{V}_{ZY}^{(n)}.
\]

(52)

E. Cross-validation procedure

Obtaining a kernel (or more precisely – a Gramm matrix) is computationally intensive. Performing cross-validation requires calculation of 2 kernels (one for testing data, the other for validation data) for each point of the grid. It is most effective to calculate one kernel for testing and validation data at the same time. This is done by ordering the data so that the training data points are subsequent (and validation point are subsequent), calculating kernel for the whole (but appropriately ordered) dataset and selecting the appropriate part of the kernel for testing and validation:

\[
\bar{K} = \begin{bmatrix}
K(W_{train}, W_{train}) & K(W_{train}, W_{val}) \\
K(W_{val}, W_{train}) & K(W_{val}, W_{val})
\end{bmatrix}
\]

(54)
The kernel for the validation point is now a part of the kernel that used both the training and validation points:

\[ \bar{K}_{\text{val}} = K(W_{\text{val}}, W_{\text{train}}) \] (55)

Such approach is important because it allows us to use the dual parameters calculated for the testing data without problems with dimensions. Recalling equation 14 we can now express the error as:

\[ \frac{1}{m} (\bar{K}_{\text{val}} \alpha^* - x)^T (\bar{K}_{\text{val}} \alpha^* - x) \] (56)

Even with an effective way of calculating kernels cross validation is still expensive. As described below, to obtain significance level for a particular measurement of causality it is necessary to calculate permutation tests and to obtain an acceptance rate or a series of p-values for a moving window. In consequence, to be capable of running many experiments and using several measures in a reasonable amount of time we needed a few sacrifices.

A possible approach is not to perform the cross-validation each time when running the experiments. Instead we would perform cross-validation once per experiment and use those parameters in all of the trials of the experiment. While this could have influenced the reported results for kernelised Geweke’s measure, it is reasonable because the parameters obtained by cross-validation are similar for data adhering to the same distribution. We believe that one of the strengths of kernelised Geweke’s measure, and one of the reasons why kernels are often used for online learning, lie in the fact that it is possible to optimise the parameters, but the parameters don’t have to be optimised each time.

The Geweke’s measures are based on the optimal linear prediction. While we generalise them to use nonlinear prediction we can still use the optimal predictor if we employ cross-validation.

In the applications described in the paper we have used use the Gaussian kernel, which is defined as follows:

\[ k(x, y) = \exp\left(-\frac{\|x - y\|^2}{\sigma^2}\right), \] (57)

and the linear kernel \( k(x, y) = x^T y \).

We use a randomized 5-fold cross-validation to choose the optimal parameter \( \gamma \) for the regularisation and the kernel parameters. Let \( (x_t, y_t, z_t), t = 1, \ldots, n \) be the time series. We want to calculate \( G_{y \to x; z} \). Based on the given time series we create a learning set and lag (embedding) equal \( p \) we prepare a learning set, following the notation from section 2.1.3. : \( (x_t, w_{t-p}^T) \), for \( i = p + 1, \ldots, n \). The learning set is split randomly into 5 subsets of equal size. For each \( k = 1, \ldots, 5 \) we obtain a k-th validation set and and a k-th testing set that contains all data points that don’t belong to the k-th validation set.

Next a grid is created, given a range of values for the parameter \( \gamma \) and for the the kernel parameters (the values change in logarithmic scale). For each training set and each point on the grid we calculate the dual weights \( \alpha^* \). Those dual weights are used to calculate the validation score – the prediction error for this particular grid point. The five validation scores are averaged to obtain an estimate of the prediction error for each of the points on the grid. We choose those parameters that correspond to the point of grid with minimum estimate of the prediction error. Finally, we calculate the prediction error on the whole learning set given the chosen optimal parameters.
As mentioned, the set of parameters from which we choose the optimal one is spread across a logarithmic scale. The whole cross-validation can be relatively expensive computationally and therefore unnecessarily big grid is undesirable.

Acknowledgements

Many thanks to Kacper Chwiałkowski (University College London) for useful discussions and valuable feedback on the manuscript. Special thanks to Maciej Makowski for proofreading and providing general comments to the initial draft. We would also like to express gratitude to Max Lungarella, CTO of Dynamic Devices AG, for making available the code of [26] which, while not used in any of the experiments described here, has provided us with important insights into transfer entropy and alternative methods.

Conflict of Interest

The authors declare no conflict of interest.

References

1. Pearl, J. Causality: models, reasoning, and inference; Cambridge University Press: New York, NY, USA, 2000.
2. Granger, C.W.J. Testing for causality : A personal viewpoint. Journal of Economic Dynamics and Control 1980, 2, 329–352.
3. Wiener, Norbert. The theory of prediction. In Modern Mathematics for Engineers; Edwin F. Beekenbach., Ed.; New York: McGraw-Hill, 1956; Vol. 1.
4. Granger, C. Economic processes involving feedback. Information and Control 1963, 6, 28–48.
5. Granger, C.W.J. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. Econometrica 1969, 37, 424–438. ArticleType: research-article / Full publication date: Aug., 1969 / Copyright 1969 The Econometric Society.
6. Geweke, J. Measurement of Linear Dependence and Feedback between Multiple Time Series. Journal of the American Statistical Association 1982, 77, 304–313.
7. Geweke, J.F. Measures of Conditional Linear Dependence and Feedback Between Time Series. Journal of the American Statistical Association 1984, 79, 907–915. ArticleType: research-article / Full publication date: Dec., 1984 / Copyright 1984 American Statistical Association.
8. Amblard, P.; Vincent, R.; Michel, O.; Richard, C. Kernelizing Geweke’s measures of granger causality. 2012 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), 2012, pp. 1–6.
9. Amblard, P.O.; Michel, O.J.J. The relation between Granger causality and directed information theory: a review. arXiv e-print 1211.3169, 2012.
10. Gretton, A.; Herbrich, R.; Smola, A.; Bousquet, O.; Schlkopf, B. Kernel Methods for Measuring Independence. J. Mach. Learn. Res. 2005, 6, 20752129.
11. Gretton, A.; Bousquet, O.; Smola, A.; Schlkopf, B. Measuring statistical dependence with Hilbert-Schmidt norms. PROCEEDINGS ALGORITHMIC LEARNING THEORY. Springer-Verlag, 2005, p. 6377.
12. Sun, X.; Janzing, D.; Schlkopf, B.; Fukumizu, K. A kernel-based causal learning algorithm. Proceedings of the 24th international conference on Machine learning; ACM: New York, NY, USA, 2007; ICML ’07, p. 855862.
13. Fukumizu, K.; Gretton, A.; Sun, X.; Schlkopf, B. Kernel Measures of Conditional Dependence, 2008.
14. Gretton, A.; Fukumizu, K.; Teo, C.H.; Song, L.; Schlkopf, B.; Smola, A.J. A kernel statistical test of independence. Technical report, 2008.
15. X. Sun.; D. Janzing.; B. Schlkopf.; K. Fukumizu. A kernel-based causal learning algorithm. Proc. 24th Intern. Conf. Machine Learning, 2007, pp. 855–862.
16. Guyon, I.; Janzing, D.; Schlkopf, B. Causality: Objectives and Assessment. 2008.
17. Sun, X. Assessing Nonlinear Granger Causality from Multivariate Time Series. In Machine Learning and Knowledge Discovery in Databases; Daelemans, W.; Goethals, B.; Morik, K., Eds.; Number 5212 in Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2008; pp. 440–455.
18. Schlkopf, E.b.B.; Vert, K.T.; Jean-Philippe. Kernel Methods in Computational Biology 2004.
19. Hastie, T.; Tibshirani, R.; Friedman, J.H. The elements of statistical learning data mining, inference, and prediction; Springer: New York, 2009.
20. Renyi, A. On measures of dependence. Acta Mathematica Academiae Scientiarum Hungarica 1959, 10, 441–451.
21. Gretton, A.; Smola, A.; Bousquet, O.; Herbrich, R.; Belitski, A.; Augath, M.; Murayama, Y.; Pauls, J.; Schoelkopf, B.; Logothetis, N. Kernel Constrained Covariance for Dependence Measurement, 2005.
22. Hofmann, T.; Schlkopf, B.; Smola, A.J. A review of RKHS methods in machine learning. Technical report, 2006.
23. Seth, S.; Principe, J. Assessing Granger Non-Causality Using Nonparametric Measure of Conditional Independence. IEEE Transactions on Neural Networks and Learning Systems 2012, 23, 47–59.
24. Schreiber, T. Measuring Information Transfer. arXiv:nlin/0001042 2000. Phys. Rev. Lett. 85 (2000) 461.
25. Lindner, M.; Vicente, R.; Priesemann, V.; Wibral, M. TRENTOOL: A Matlab open source toolbox to analyse information flow in time series data with transfer entropy. BMC Neuroscience 2011, 12, 119. PMID: 22098775.
26. Lungarella, M.; Ishiguro, K.; Kuniyoshi, Y.; Otsu, N. METHODS FOR QUANTIFYING THE CAUSAL STRUCTURE OF BIVARIATE TIME SERIES. International Journal of Bifurcation and Chaos 2007, 17, 903–921.
27. Amblard, P.O.; Michel, O.J.J. On directed information theory and Granger causality graphs. Journal of Computational Neuroscience 2011, 30, 7–16.
28. Barnett, L.; Barrett, A.B.; Seth, A.K. Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables. *Physical Review Letters* **2009**, *103*, 238701.
29. Seth, A.K. A MATLAB toolbox for Granger causal connectivity analysis. *Journal of neuroscience methods* **2010**, *186*, 262–273. PMID: 19961876.
30. Hlavackova-Schindler, K.; Palus, M.; Vejmelka, M.; Bhattacharya, J. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports* **2007**, *441*, 1–46.
31. Chvez, M.; Martinerie, J.; Le Van Quyen, M. Statistical assessment of nonlinear causality: application to epileptic EEG signals. *Journal of neuroscience methods* **2003**, *124*, 113–128. PMID: 12706841.
32. Knuth, K.H.; Golera, A.; Curry, C.T.; Huyser, K.A.; Wheeler, K.R.; Rossow, W.B. Revealing Relationships among Relevant Climate Variables with Information Theory **2005**.
33. Gourvitch, B.; Eggermont, J.J. Evaluating information transfer between auditory cortical neurons. *Journal of neurophysiology* **2007**, *97*, 2533–2543. PMID: 17202243.
34. Vicente, R.; Wirbal, M.; Lindner, M.; Pipa, G. Transfer entropy—a model-free measure of effective connectivity for the neurosciences. *Journal of computational neuroscience* **2011**, *30*, 45–67. PMID: 20706781.
35. Bressler, S.L.; Seth, A.K. Wiener-Granger causality: a well established methodology. *NeuroImage* **2011**, *58*, 323–329. PMID: 20202481.
36. Alexander, C.; Wyeth, J. Cointegration and market integration: An application to the Indonesian rice market. *Journal of Development Studies* **1994**, *30*, 303–334.
37. Cont, R. Long range dependence in financial markets. Fractals in Engineering. Springer, 2005, p. 159180.
38. Patton, A.J. CopulaBased Models for Financial Time Series. In *Handbook of Financial Time Series*; Mikosch, T.; Krei, J.P.; Davis, R.A.; Andersen, T.G., Eds.; Springer Berlin Heidelberg, 2009; pp. 767–785.
39. Durante, F. Copulas, Tail Dependence and Applications to the Analysis of Financial Time Series. In *Aggregation Functions in Theory and in Practise*; Bustince, H.; Fernandez, J.; Mesar, R.; Calvo, T., Eds.; Number 228 in Advances in Intelligent Systems and Computing, Springer Berlin Heidelberg, 2013; pp. 17–22.
40. Barber, D. *Bayesian Reasoning and Machine Learning*; Cambridge University Press, 2012.
41. Eichler, M. Granger causality and path diagrams for multivariate time series. *Journal of Econometrics* **2007**, pp. 334–353.
42. An Pan.; Qi Sun.; Adam M. Bernstein.; Matthias B. Schulze.; JoAnn E. Manson.; Meir J. Stampfer.; Walter C. Willett.; Frank B. Hu. Red meat consumption and mortality: Results from 2 prospective cohort studies. *Archives of Internal Medicine* **2012**, *172*, 555–563.
43. Friedman, M. The Fed’s Thermostat. *Wall Street Journal* **2003**.
44. Fenton, N.; Neil, M. *Risk Assessment and Decision Analysis with Bayesian Networks*; CRC Press, 2012.
45. Hsiao, C. Autoregressive modelling and causal ordering of economic variables. *Journal of Economic Dynamic and Control* **1982**, p. 243259.
46. Papana, A.; Kyrtsou, C.; Kugiumtzis, D.; Diks, C. Simulation Study of Direct Causality Measures in Multivariate Time Series. *Entropy* 2013, 15, 2635–2661.

47. Kugiumtzis, D. Partial transfer entropy on rank vectors. *The European Physical Journal Special Topics* 2013, 222, 401–420.

48. Lungarella, M.; Pitti, A.; Kuniyoshi, Y. Information transfer at multiple scales. *Physical review. E, Statistical, nonlinear, and soft matter physics* 2007, 76, 056117. PMID: 18233728.

49. Dhamala, M.; Rangarajan, G.; Ding, M. Estimating Granger causality from Fourier and wavelet transforms of time series data. *Physical review letters* 2008, 100, 018701. PMID: 18232831.

50. Kenji Fukumizu. Kernel Methods for Dependence and Causality. *Machine Learning Summer School* 2007.

51. Shawe-Taylor, J.; Cristianini, N. *Kernel methods for pattern analysis*; Cambridge University Press: New York, 2004.

52. Hatemi-J, A. Asymmetric causality tests with an application. *Empirical Economics* 2012, 43, 447–456.

53. Dahlhaus, R.; Eichler, M. Causality and graphical models in time series analysis. In *Highly structured stochastic systems*; Green, P.J.; Hjort, N.L.; Richardson, S., Eds.; Oxford University Press: Oxford; New York, 2003.

54. Steinwart, I.; Christmann, A. *Support vector machines*; Springer: New York, 2008.

©