OPEN ACCESS

entropy

Article

Measures of Causality in Complex Datasets with Application to Financial Data

Anna Zaremba [1,*] **and Tomaso Aste** [1,2]

[1] Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK
[2] Systemic Risk Centre, London School of Economics and Political Sciences, London WC2A2AE, UK

* Author to whom correspondence should be addressed; E-Mail: A.Zaremba@cs.ucl.ac.uk;
Tel.: +44-0-20-3108-1305; Fax: +44-0-20-7387-1397.

Abstract: This article investigates the 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 the 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 present the theoretical benefits of applying non-symmetrical measures rather than symmetrical measures of dependence. We apply the measures to a range of simulated and real data. The simulated data sets were generated with linear and several types of nonlinear dependence, using bivariate, as well as multivariate settings. An application to real-world financial data highlights the practical difficulties, as well as the potential of the methods. We use two real data sets: (1) U.S. inflation and one-month Libor; (2) S&P data and exchange rates for the following currencies: AUDJPY, CADJPY, NZDJPY, AUDCHF, CADCHF, NZDCHF. Overall, we reach 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 highlight areas for improvement and future research.

1. Introduction

Understanding the dependence between time series is crucial for virtually all complex systems studies, and the ability to describe the 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 one to answer questions, like "if we give a drug to a patient, *i.e.*, intervene, will their chances of survival increase?". Statistical causality does not answer such questions, because it does not operate on the concept of intervention and only involves tools of data analysis. Therefore, the causality in a 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 say 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 discernible 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 do not follow his nomenclature, we agree that it is 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 the lack of stationarity and the presence of noise. If we also consider that the dependence is likely to exhibit nonlinearity, model selection becomes an important factor that needs to be better understood.

The goal of this paper is to provide a broad analysis of several of the existing methods to quantify 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 the practical aspects, the details of implementation, as well as the methodology of testing and the results of testing on synthetic data; financial and other application are described in Section 4. In Section 5, we provide a discussion of the methods, applications and perspectives. Section 6 contains a brief summary. Finally, in Appendices A–E, we provide the mathematical background and other supplementary material.

2. Methodology: Literature Review

2.1. Definitions of Causality, Methods

The first mention of causality as a property that can be estimated 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 1963 [4] and 1969 [5]. The context in which Granger defined causality was that of linear autoregressive models of stochastic processes. Granger 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 discussions of what causality means and how the statistical concept he introduced differed from deterministic causation.

2.1.1. Granger Causality

In the 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 is Granger causality if the notion of time is introduced and 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 [8].

We use notation and definitions that derive from [9], and we generalise them. Let $\{X_t\}, \{Y_t\}, \{Z_t\}$ be three stochastic processes. For any of the time series, using subscript t, as in X_t, it will be understood as a random variable associated with the time t, while using superscript t, as in X^t, it will be understood as the collection of random variables up to time t. Accordingly, we 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 $t \in \mathbb{Z}$:*

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

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

$$P\left(X_t \mid X^{t-1}, Y^t, Z^{t-1}\right) = P\left(X_t \mid X^{t-1}, Z^{t-1}\right). \tag{2}$$

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 it is only one of several alternative definitions,

none of which is *a priori* optimal. Amblard [10] recommends including Z^t rather than Z^{t-1} to ensure that the measure precludes confusing [11] the instantaneous coupling of X and Y with that of X and Z. Definition 1 is very general and does not 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 [9].

Let us here start by introducing the measures of (Granger) causality that were originally proposed by Geweke in [6,7]. 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_X, L_{YX}, L_{ZX} are linear functions. In Equation (3), we allow some of the functions, $L_\bullet$, 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 $L_{YX} = 0$:

$$X_t = \tilde{L}_X(X^{t-1}) + \tilde{L}_{ZX}(Z^{t-1}) + \tilde{\varepsilon}_{X,t} \tag{4}$$

then we can quantify the usefulness of including $\{Y_t\}$ in explaining $\{X_t\}$ with Geweke's measure of causality formulated as follows:

$$F_{Y \to X \| Z} = \log \frac{var(X_t \mid X^{t-1}, Z^{t-1})}{var(X_t \mid X^{t-1}, Y^{t-1}, Z^{t-1})} = \log \frac{var(\tilde{\varepsilon}_{X,t})}{var(\varepsilon_{X,t})}. \tag{5}$$

Analogously, Geweke's measure of instantaneous causality (instantaneous coupling) is defined as:

$$F_{Y \cdot X \| Z} = \log \frac{var(X_t \mid X^{t-1}, Z^{t-1})}{var(X_t \mid X^{t-1}, Y^t, Z^{t-1})} = \log \frac{var(\tilde{\varepsilon}_{X,t})}{var(\hat{\varepsilon}_{X,t})}, \tag{6}$$

where $\hat{\varepsilon}_{X,t}$ corresponds to yet another possible model, where both the past and present of Y are 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 a later section, we will present a generalisation of Geweke's measure using kernel methods.

2.1.2. Kernels

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

There is a rich body of literature on causal inference from the machine learning perspective. Initially, the interest concentrated on testing for independence [12–16]; 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 [14,17].

In particular, in the last several years, kernelisation has become a popular approach for generalising linear algorithms in many fields. The main idea underlying 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 and searching for a 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. As a function of two variables, the kernel function can be interpreted as a comparison function. It can also be thought 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, Amblard *et al.* published the paper [9], which, to the best of our knowledge, is the first suggesting the 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 a straightforward way of optimising parameters.

An alternative approach to kernelising Granger causality was proposed by Xiaohai Sun in [18]. Sun proposed the use of the square root of the Hilber–Schmidt norm of the so-called conditional cross-covariance operator (see Definition 3) in the feature space to measure the prediction error and the use of a permutation test to quantify the improvement of predictability. While neither of the two kernel approaches described in this paper is based on Sun's article, they are closely related. In particular, the concept of the Hilbert–Schmidt Normalised Conditional Independence Criterion (HSNCIC) is from a similar family as the one explored by Sun. Another method of kernelising Granger causality has been described by Marinazzo *et al.* [19]. Below, we are following the approach from [20]. Please refer to the Appendix B for supplementary information on functional analysis and Hilbert spaces.

Let us denote with $S = (x_1, ..., x_n)$ 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 $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ of the points of the set, S, rather that 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 $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a positive definite kernel if and only if it is symmetric, that is, $\forall x, x' \in \mathcal{X}, k(x, x') = k(x', x)$ and positive (semi-) definite, that is:*

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

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

Theorem 1 *For any kernel, k, on space $\mathcal{X}$, there exists a Hilbert space, F, and a mapping $\phi : \mathcal{X} \to F$, such that [20]:*

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

where $\langle u, v \rangle, u, v \in \mathcal{F}$ represents an inner product in $\mathcal{F}$.

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 an identity map, this theorem proves that the 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. The kernel trick can be stated as follows [20]: "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 the nonlinear case of linear regression-Granger causality; and (2) the reformulation of concepts, such as covariance and partial correlations, to the 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. To do so, 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 now assume to have a set of observations $S = \{(x_1, y_1, z_1), ..., (x_n, y_n, z_n)\}$, $x_t, y_t \in \mathbb{R}, z_t \in \mathbb{R}^k$. The goal is to find the best linear fit function, f, which describes the given dataset, S. The idea is similar to the linear Geweke's measure: to infer causality from the comparison of alternative models. In particular, four alternative functional relations between points are modelled: (1) between x_t and its own past; (2) between x_t and the past of x_t, y_t; (3) between x_t and the past of x_t, z_t; and (4) between x_t and the past of x_t, y_t, z_t. In order to have a uniform notation for all four models, a new variable, w, is introduced, with w_i symbolising either x_i itself, or (x_i, y_i), or (x_i, z_i) or (x_i, y_i, z_i). Thus, the functional relationship between the data can be written as follows: for all t, $x_t \simeq f(w_{t-1}^{t-p})$, where 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, p could represent 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 $w = x$, or $d = 2p$ if $w = (x, y)$ or $d = 2p + kp$ if $w = (x, y, z)$. Least squares regression (as in the linear Granger causality discussed earlier) involves looking for a real valued weight vector, β, such that $x_t \simeq \hat{x}_t = (w_{t-1}^{t-p})^T \beta$, *i.e.*, choosing the weight vector, β, that minimises the squared error. The dimensionality of β depends on the dimensionality of w; it is a scalar in the simplest case of $w = x$, with x being univariate.

It is well known that the drawbacks of least squares regression include: poor effects with small sample size, no solution when data are linearly independent and overfitting. Those problems can be addressed by adding to the cost function an additional cost penalizing excessively large weights of the coefficients. This cost, called the regulariser [21] or regularisation term, introduces a trade-off between the mean squared error and a squared norm of the weight vector. The regularised cost function is now:

$$\beta^* = \underset{\beta}{\operatorname{argmin}} \frac{1}{m} \sum_{i=p+1}^{n} ((w_{i-1}^{i-p})^T \beta - x_i)^2 + \gamma \beta^T \beta, \tag{9}$$

with $m = n - p$ for a more concise notation.

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

$$\beta^* = (\mathbf{W}^T\mathbf{W} + \gamma m I_m)^{-1}\mathbf{W}^T x, \tag{10}$$

where we use the matrix notation $\mathbf{W} = ((w_p^1)^T, (w_{p+1}^2)^T, ..., (w_{n-1}^{n-p})^T)^T$, or in other words, a matrix with rows $w_p^1, w_{p+1}^2, ..., w_{n-1}^{n-p}$; $x = (x_{p+1}, x_{2,t}, ..., x_n)^T$; and 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, β, can be represented as a linear combination of the data points: $\beta = \mathbf{W}^T\alpha$, for some α. This second representation results in the dual solution, α^*, that can be written in terms of $\mathbf{W}\mathbf{W}^T$ and that depends on the regulariser, γ:

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

This is where we can apply the kernel trick that will allow us to introduce kernels to the regression setting above. For this purpose, we introduce kernel similarity function k, which we apply to elements of $\mathbf{W}$. The Gram matrix built from evaluations of kernel functions on each row of $\mathbf{W}$ is denoted by $\mathbf{K_w}$:

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

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

$$\hat{x}_t = k_w(w_{t-1}^{t-p})^T(\mathbf{K_w} + \gamma m I_m)^{-1}x. \tag{13}$$

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

$$var_K(X_t \mid W^{t-1}) = \frac{1}{m}\sum_{j=1}^l (x_j - \hat{x}_j)^2 = \frac{1}{m}(\mathbf{K_w}\alpha^* - x)^T(\mathbf{K_w}\alpha^* - x), \tag{14}$$

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

Analogously to the Geweke's indices from Equations (5), we now define kernelised Geweke's indices for causality and instantaneous coupling using the above framework:

$$\begin{aligned}
G_{Y \to X \| Z} &= \log \frac{var_K(X_t \mid X^{t-1}, Z^{t-1})}{var_K(X_t \mid X^{t-1}, Y^{t-1}, Z^{t-1})} \\
G_{Y \cdot X \| Z} &= \log \frac{var_K(X_t \mid X^{t-1}, Z^{t-1})}{var_K(X_t \mid X^{t-1}, Y^t, Z^{t-1})},
\end{aligned} \tag{15}$$

extending in this way Geweke's measure of causality to the non-linear case.

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 [22] pointed out

that to assess independence between random variables X and Y, one can use maximum correlation S defined as follows:

$$S(X, Y) = \sup_{f,g} \left(corr(f(X), g(Y)) \right) \qquad (16)$$

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 that Renyi postulated for an appropriate measure of dependence; most importantly, it equals zero if and only if the variables, X and Y, are independent. However, the concept of maximum correlation is not practical, as there might not even exist such functions, f_0 and g_0, for which the maximum can be attained [22]. Nevertheless, this concept has been used as a foundation of some kernel-based methods for dependence, such as kernel constrained covariance [23].

This section requires some background from functional analysis and machine learning. For completeness, the definitions of the Hilbert–Schmidt norm and operator, tensor product and mean element are given in the Appendix B and follow [13,15].

The cross-covariance operator is analogous to a covariance matrix, but is defined for feature maps.

Definition 3 *(Cross-covariance operator)* *The cross-covariance operator is a linear operator* Σ_{XY} : $\mathcal{H}_{\mathcal{Y}} \to \mathcal{H}_{\mathcal{X}}$ *associated with the joint measure,* P_{XY}, *defined as:*

$$\Sigma_{XY} := \boldsymbol{E}_{XY}[(\phi(X) - \mu_X) \otimes (\phi(Y) - \mu_Y)] = \boldsymbol{E}_{XY}[\phi(X) \otimes \phi(Y)] - \mu_X \otimes \mu_Y \qquad (17)$$

where we use symbol $\otimes$ for tensor product and μ for mean embedding (definitions in Appendix B). The cross-covariance operator applied to two elements of $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ gives the covariance:

$$\langle f, \Sigma_{XY} g \rangle_{\mathcal{H}_{\mathcal{X}}} = Cov(f(X), g(Y)) \qquad (18)$$

The notation and assumptions follow [13,18]: $\mathcal{H}_{\mathcal{X}}$ denotes the reproducing kernel Hilbert space (RKHS) induced by a strictly positive kernel $k_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, analogously for $\mathcal{H}_{\mathcal{Y}}$ and $k_{\mathcal{Y}}$. 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 $\mathcal{Y}$ to be topological spaces, and measurability is defined with respect to the relevant $\sigma-$fields. The marginal distributions are denoted by P_X, P_Y and the joint distribution of (X, Y) by P_{XY}. The expectations, $\mathbf{E}_X$, $\mathbf{E}_Y$ and $\mathbf{E}_{XY}$, denote the expectations over P_X, P_Y and P_{XY}, respectively. To ensure $\mathcal{H}_{\mathcal{X}}, \mathcal{H}_{\mathcal{Y}}$ are included in, respectively, $L^2(P_X)$ and $L^2(P_Y)$, we consider only random vectors (X, Y), such that the expectations, $\mathbf{E}_X[k_{\mathcal{X}}(X, X)]$ and $\mathbf{E}_Y[k_{\mathcal{Y}}(Y, Y)]$, are finite.

Just as the 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 [15]) Using the cross-covariance operators, we can define the normalised conditional cross-covariance operator in the following way:*

$$V_{XY|Z} = \Sigma_{XX}^{-1/2} (\Sigma_{XY} - \Sigma_{XZ} \Sigma_{ZZ}^{-1/2} \Sigma_{ZY}) \Sigma_{YY}^{-1/2} \qquad (19)$$

Gretton *et al.* [13] state that for rich enough RKHS (by "rich enough", we mean universal, *i.e.*, dense in the sense of continuous functions on $\mathcal{X}$ with the supremum norm [24]), the zero norm of the cross-covariance operator is equivalent to independence, which can be written as:

$$X \perp\!\!\!\perp Y \Longleftrightarrow \Sigma_{XY} = 0 \tag{20}$$

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

It was shown in [15] that there is a relationship similar to (20) between the normalised conditional cross-covariance operator and conditional independence, which can be written as:

$$X \perp\!\!\!\perp Y \mid Z \Longleftrightarrow V_{(XZ)(YZ)|Z} = 0 \tag{21}$$

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 the non-normalised operator, while retaining all the information about dependence. This is analogous 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, $V_{(XZ)(YZ)|Z}$:*

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

where $\|\cdot\|_{HS}$ denotes Hilbert–Schmidt norm of an operator, defined in the Appendix B.

For the sample $S = \{(x_1, y_1, z_1), ..., (x_n, y_n, z_n)\}$, HSNCIC has an estimator that is both straightforward and has good convergence behaviour [15,25]. As shown in 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 V_{XY} requires inverting Σ_{YY} and Σ_{XX}; hence, to ensure invertibility a regulariser, $n\lambda I_n$, is added. The next step is to construct the estimator, $\hat{V}_{XY|Z}^{(n)}$, from $\hat{V}_{XY}^{(n)}$, $\hat{V}_{XZ}^{(n)}$ and $\hat{V}_{ZY}^{(n)}$. Finally, construct the estimator of the Hilbert–Schmidt norm of $\hat{V}_{ZY}^{(n)}$ as follows:

$$HSNCIC_n := Tr[R_{(XZ)}R_{(YZ)} - 2R_{(XZ)}R_{(YZ)}R_Z + R_{(XZ)}R_Z R_{(YZ)}R_Z] \tag{23}$$

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, λ, which, in turn, depends on the sample size. Regularisation becomes necessary when inverting finite rank operators.

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" [26]. Transfer entropy has become popular

among physicists and biologists, and there is a large body of literature on the application of transfer entropy to neuroscience. We refer to [27] for a description of one of the best developed toolboxes for estimating transfer entropy. A comparison of transfer entropy and other methods to measure causality in bivariate time series, including extended Granger causality, nonlinear Granger causality, predictability improvement and two similarity indices, was presented by Max Lungarella *et al.* in [28]. A particularly exhaustive review of the relation between Granger causality and directed information is presented by Amblard *et al.* [10], while for a treatment of the topic from the network theory perspective, refer to Amblard and Michel [29].

Transfer entropy was designed to measure the departure from the generalized Markov property stating that $P\left(X_t \mid X^{t-1}, Y^{t-1}\right) = P\left(X_t \mid X^{t-1}\right)$. 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.* [30] for the 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 $P_U(u_i), P_V(v_i)$ and joint distribution $P_{UV}(u_i, v_i)$. Then, the mutual information, $I(U, V)$, is defined as:*

$$I(U, V) = \sum_{i,j} P_U(u_i, v_j) \log \frac{P_{UV}(u_i, v_j)}{P_U(u_i) P_V(v_j)} = H(U) - H(U \mid V) \tag{24}$$

with $H(U) = -\sum_i P_U(u_i) \log P_U(u_i)$ the Shannon entropy and $H(U|V) = \sum_{i,j} P_{UV}(u_i, v_j) \log \frac{P_V(v_j)}{P_{UV}(u_i, v_j)}$ the Shannon conditional entropy.

For independent random variables, the 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. However, being a symmetrical measure, mutual information cannot provide any information about the 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 distributions, while transfer entropy is a generalisation to more than one system and is defined in terms of transition probabilities [26].

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 $T_{Y \to X}$ is defined as:*

$$T_{Y \to X} = H(X_t \mid X^{t-1}) - H(X_t \mid X^{t-1}, Y^{t-1}) \tag{25}$$

Transfer entropy can be generalised for a multivariate system, for example [30] defines conditional transfer entropy $T_{Y \to X|Z} = H(X_t \mid X^t, Z^t) - H(X_t \mid X^t, Y^t, Z^t)$. In this paper, we will calculate transfer entropy only in the case of two variables. This is because the calculations already involve the estimation of the joint distribution of three variables (X_t, X^t, Y^t), and estimating the joint

distribution of more variables would be impractical for time series of the length that we work with in financial applications.

3. Testing

3.1. Permutation Tests

Let us, first of all, emphasise that in the general case, the causality measures introduced before 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 the 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 the significance of the causal relationship itself. To achieve this goal, we shall use permutation tests, following the approach in [9,18,25].

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. We would like to compare the value of our causality measure on the analysed data and on "random" data and conclude that the former is significantly higher. 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 $\pi_1, ..., \pi_{n_r}$ be a set of random permutations. Then, instead of y_t, we consider $y_{\pi_j(t)}$, obtaining a set of measurements $G_{Y_{\pi_j} \to X || Z}$ that can be used as an estimator of the null hypothesis $G_{Y \to X || Z}^0$. 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*-value defined as follows:

$$p = \frac{1}{n_r} \sum_{j=1}^{n_r} \mathbf{1}(G_{Y_{\pi_j} \to X || Z} > G_{Y \to X || Z}) \tag{26}$$

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 report 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 do not believe that using any kind of subsampling (as proposed by [9,18,25]) will be beneficial, as far as the power of the tests is concerned.

3.2. Testing on Simulated Data

3.2.1. Linear Bivariate Example

Before applying the methods to real-world data, it is prudent to verify whether they work 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 was 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 were shifted by one, two or three time steps to obtain the following "causal" relations: $x_1 \longleftrightarrow x_2$ 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 Figure 1, while the lags at which the causality occurs are given in the Table 1b. The length of the data is 250.

Table 1. The dependence structure of the simulated data. (**a**) The correlation matrix that has been used to generate the testing data; (**b**) Lags at which true dependence occurs, with the interpretation that the column variable causes the row variable.

	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

(a)

	ts1	ts2	ts3	ts4	ts5	ts6	ts7	ts8
ts1	×	0						
ts2	0	×						
ts3			×	-1				
ts4			1	×				
ts5					×	-1	-2	-3
ts6					1	×	-1	-2
ts7					2	1	×	-1
ts8					3	2	1	×

(b)

Figure 1. The directionality of causality between the eight simulated time series. Green lines represent causality, with the arrowheads indicating direction; the red line indicates instantaneous coupling.
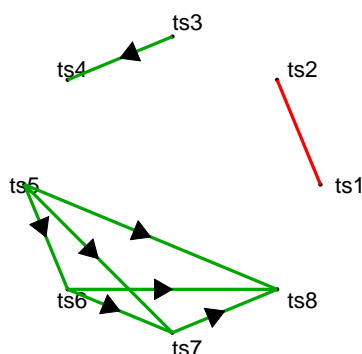
For the purpose of the experiments described in this paper, we used code from several sources: Matlab code that we developed for kernelised Geweke's measure and transfer entropy, open access Matlab toolbox for Granger causality GCCA [31,32] and open access Matlab code provided by Sohan Seth [25,33].

To calculate Geweke's measure and kernelised Geweke's measure, we used the same code, with a linear kernel in the former case and a Gaussian kernel in the latter. The effect of regularisation on the (linear) Geweke's measure is negligible, and the results are comparable to the ones obtained with 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 (see Appendix E) for the grid of regulariser 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 preset level, with no noticeable impact on the result. Transfer entropy utilises a naive histogram to estimate distributions. The code for calculating HSNCIC and for performing *p*-value tests incorporates a framework written by Seth [25]. The framework has been altered to accommodate some new functionalities; the implementation of permutation tests has also been changed from rotation to actual permutation [34] In the choice of parameters for the HSNCIC, we followed [25], where the size of the kernel is set up as the median inter-sample distance and regularisation is set to 10^{-3}.

Our goal was to uncover the causal structure without prior information and detect the lags at which causality occurred. This was performed by applying all three measures of causality with the following sets of lags: $\{[1-10]\}, \{[1-20]\}, \{[1-5], [6-10], [11-15]\}, \{[1-3], [4-6], [7-9]\}$; and finally, with all four measures to single lags $\{0, 1, 2, 3, 4\}$. Those ranges were used for linear and kernelised Geweke's measures and HSNCIC, but not for transfer entropy, for which only single lags are available with the current framework. Using five sets of lags allowed us to analyse the effects of using ranges of lags that are different from lags corresponding to the "true" dynamic of the variables. 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 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 zero 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 Lag 0, *i.e.*, instantaneous coupling. However, with the number of permutations equal 200 and at an acceptance level of 0.01, the two measures detected only the required causalities, but would fail to reject some spurious causalities at a 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. In order to detect this, we applied the mutual information method instead. For the lags $\{1, 2, 3\}$, the transfer entropy reported zero *p*-values for all the relevant causal directions. However, it failed to reject the spurious direction $1 \rightarrow 7$ with a *p*-value of 0.01. For lag $\{0\}$, where mutual information has been applied, the instantaneous coupling $x_1 \longleftrightarrow x_2$ was recognised correctly with a *p*-value of zero.

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 zero. This 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.

Table 2. *p*-values for four measures for Lag 1. From top left to bottom right: Geweke's measure (Gc), kernelised Geweke's measure (kG), transfer entropy (TE), HSNCIC (HS). All Lag 1 causalities were correctly retrieved by all methods.

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**	×

TE	ts1	ts2	ts3	ts4	ts5	ts6	ts7	ts8
ts1	×	0.59	0.53	0.60	0.34	0.91	0.38	0.66
ts2	0.48	×	0.86	0.30	0.87	0.96	0.49	0.70
ts3	0.45	0.17	×	0.33	0.34	0.57	0.81	0.81
ts4	0.04	0.31	**0**	×	0.12	0.09	0.76	0.08
ts5	0.21	0.52	0.86	0.05	×	0.68	0.60	0.30
ts6	0.53	0.89	0.65	0.30	**0**	×	0.77	0.09
ts7	0.01	0.42	0.59	0.37	0.95	**0**	×	0.77
ts8	0.85	0.46	0.07	0.48	0.85	0.13	**0**	×

HS	ts1	ts2	ts3	ts4	ts5	ts6	ts7	ts8
ts1	×	1.00	0.81	0.35	0.19	0.48	0.71	0.82
ts2	1.00	×	0.80	0.61	0.85	0.34	0.02	0.72
ts3	0.90	0.95	×	0.18	0.59	0.47	0.21	0.19
ts4	0.90	0.29	**0**	×	0.31	0.81	0.26	0.31
ts5	0.75	0.59	0.77	0.14	×	0.71	0.85	0.46
ts6	0.64	0.88	0.75	0.79	**0**	×	0.71	0.79
ts7	0.38	0.13	0.75	0.24	0.75	**0**	×	0.60
ts8	0.90	0.55	0.46	0.73	0.78	0.78	**0**	×

From this experiment, we conclude that Geweke's measures with linear and Gaussian kernels provide the best performance, are not vulnerable to lag misspecification and seem the most practical. The other two measures, transfer entropy and HSNCIC, provide good performance when analysing one lag at a time. In Section 3.2.2. we show the results of one of the tests from [9], which investigates the ability to distinguish between direct and non-direct causality in data where both linear and non-linear dependence have been introduced. We refer to [25] for the results of a wide range of tests applied to linear Granger causality and HSNCIC. We tested all four methods and managed to reproduced the results from [25] to a large degree; however, we used smaller number of permutations and realisations, and we obtained somewhat lower acceptance rates for true causal directions, particularly for HSNCIC. From all of those tests, we conclude 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 eight time series data described earlier, the size of the kernel did not play an important role, but in some cases, the size of the kernel was crucial in allowing the detection of causality. However, there was no kernel size that worked for all of the types of the data.

3.2.2. Nonlinear Multivariate Example

Our second example follows one presented by Amblard [9] and involves a system with both linear and non-linear causality. Apart from presenting the benefits of generalising Granger causality, this example demonstrates the potential effect of considering side information on distinguishing direct and indirect cause. The true dynamic of the time series is as follows:

$$\begin{cases} x_t = ax_{t-1} + \epsilon_{x,t} \\ y_t = by_{t-1} + dx_{t-1}^2 + \epsilon_{y,t} \\ z_t = cz_{t-1} + ey_{t-1} + \epsilon_{z,t} \end{cases} \tag{27}$$

where the parameters were chosen in the following way: $a = 0.2, b = 0.5, c = 0.8, d = 0.8, e = 0.7$, the variables $\epsilon_{x,t}, \epsilon_{y,t}, \epsilon_{z,t}$ are i.i.d. Gaussian with zero mean and unit variance. From the setup, we know that we have the following causal chain $x \rightarrow y \rightarrow z$ (with the nonlinear effect of x on y), and therefore, there is an indirect causality $x \rightarrow z$. We calculate kernelised Geweke measures $G_{x \rightarrow z}$ and $G_{x \rightarrow z|y}$ to assess the causality.

We repeat the experiment 500 times, each time generating a time series of length 500. We choose an embedding of two, *i.e.*, we consider the lag range $[1 - -2]$. To evaluate the effect of using kernelised rather than linear Granger causality, we run each experiment for the Gaussian kernel and for linear kernel $k(x, y) = x^T y$. Using the linear kernel is nearly equivalent to using the linear Geweke measures. We obtain a set of 500 measurements for $G_{x \rightarrow z}$ and $G_{x \rightarrow z|y}$, each run with a Gaussian and with a linear kernel. The results are shown in Figures 2–4. As expected, $G_{x \rightarrow z|y}$ does not detect any causality, regardless of the kernel chosen. When no side information is taken into consideration, we should see the indirect causality $x \rightarrow z$ being picked up; however, this is the case only for kernelised Geweke with the Gaussian kernel and for HSNCIC. As the dependence was nonlinear, the linear Geweke's measure did not detect it.

Figure 2. Histogram of the measurements $G_{x \rightarrow z}$ (red face), $G_{x \rightarrow z|y}$ (blue face), calculated with the kernelised Geweke's using the linear kernel (*i.e.*, the equivalent of Granger causality).
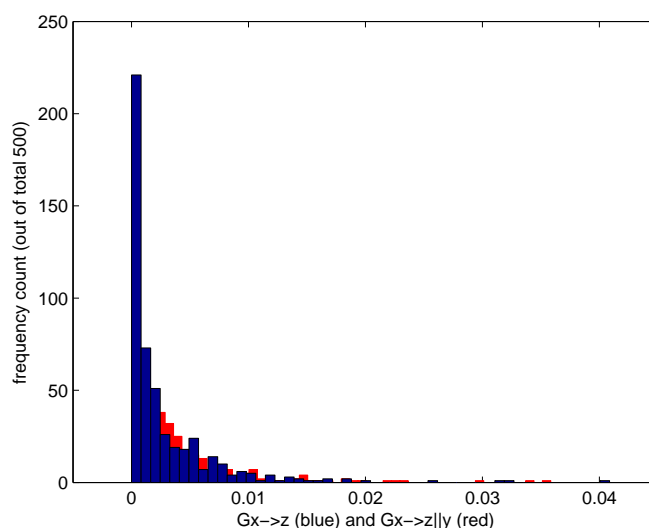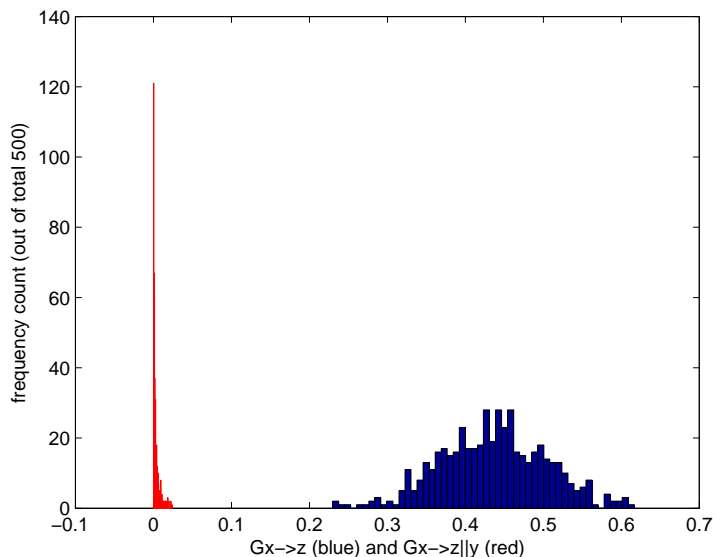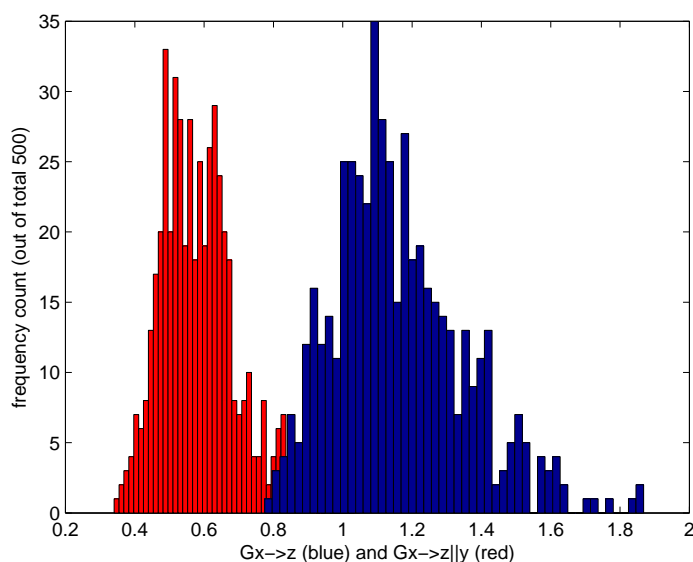
Figure 3. Histogram of the measurements $G_{x \to z}$ (red face), $G_{x \to z|y}$ (blue face), calculated with the kernelised Geweke's using the Gaussian kernel.

Figure 4. Histogram of the measurements $G_{x \to z}$ (red face), $G_{x \to z|y}$ (blue face), calculated with the Hilbert–Schmidt Normalised Conditional Independence Criterion (HSNCIC).

Transfer entropy, as defined in this paper, does not allow side information, and therefore, the result we achieve is a distribution that appears significantly different from zero (Figure 5).

Figure 5. Histogram of the measurements $G_{x \to z}$ (red face), $G_{x \to z|y}$ (blue face), calculated with the transfer entropy.
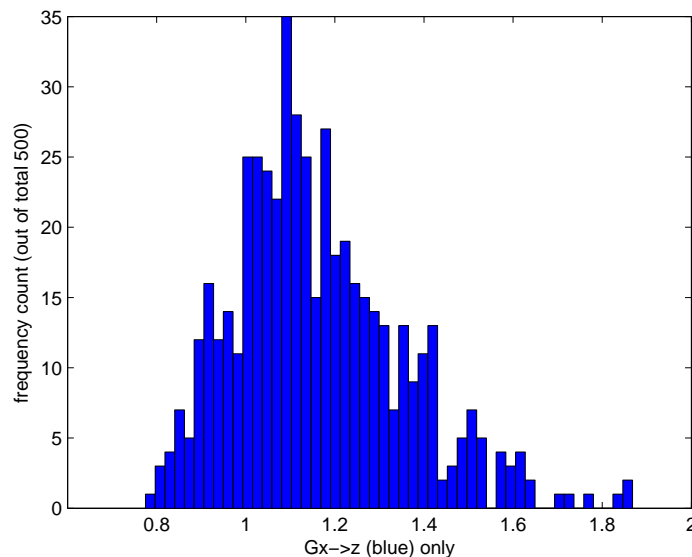
4. Applications

Granger causality was introduced as an econometrics 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 [29,35–39]. The methodology was successfully applied in those fields, particularly in neuroscience, due to the characteristics of the data common in those fields and the fact that the assumptions of Gaussian distribution and/or linear dependence are often reasonable [40]. This is generally not the case for financial time series.

4.1. Applications to Finance and Economics

In finance and economics, there are many tools devoted to modelling dependence, mostly for symmetrical dependence, such as correlation/covariance, cointegration, copula and, to a lesser degree, mutual information [41–44]. However, in various situations where we would like to reduce the dimensionality of a problem (e.g., choose a subset of instruments to invest in, choose a subset of variable for a factor model, *etc.*), knowledge of the causality structure can help in choosing the most relevant dimensions. Furthermore, forecasting using the causal time series (or Bayesian priors in Bayesian models or parents in graphical models [1,45]) helps to forecast "future rather than the past".

Financial data often have 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 to be independent identically distributed (i.i.d.). The assumption of linear dependencies or Gaussian distributions, often

argued to be reasonable in disciplines, such as neuroscience, are commonly thought to be invalid for financial time series. Furthermore, many researchers point out that stationarity usually does not apply to this kind of data. As causality methods in most cases 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 the relationship between inflation and nominal or real interest rates for the same country or region, some utilising tools of Granger causality (for example, [46]).

In this experiment, we analyse related values, namely the consumer price index for the United States (U.S. 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 is an important economic indicator. It is not a monetary measure associated with any country, and it does not reflect any institutional mandate in contrast to, for example, when the Federal Reserve sets interest rates. Instead, it reflects some level of assessment of risk by the banks who set the rate. Therefore, we ask whether we detect that one of these two economic indicators causes the other one in a statistical sense?

We ran our analysis for monthly data from January 31, 1986, to October 31, 2013, obtained from Thomson Reuters. The implementation and parameter values used for this analysis were similar to those 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 preset level in the range of $[2^7, \cdots, 2^{13}]$ or as a median.

We investigated time-windows of size 25, 50, 100 and 250. The most statistically significant and interpretable results were observed for the longer windows (250 points), where Geweke's measure and kernelised Geweke's measure show a clear indication of the direction U.S. 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 day window were seen most strongly for Lag 1 (*i.e.*, one month) and less strongly for Lags 2, 7, 8, 9, but there is no clear direction for the interim lags. In Figures 6–9, we report *p*-values for the 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 the values of causality, the lower their corresponding *p*-values.

In Figure 6, we observe that the U.S. CPI time series lagged by one month causes one-month Libor in a statistical sense, when assessed with kernelised Geweke's measure with Gaussian kernel. The *p*-values for the hypothesis of causality in this direction allow us to accept (not reject) this hypothesis at a 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 an improvement of roughly 0.18 in the explanatory power of the model [47]. Applying the linear kernel (Figure 7) resulted in somewhat similar patterns of measures of causality and *p*-values, but the two directions were less separated. Interest rates causing Libor still have *p*-values at zero most of the time, but the other direction has *p*-values that fall below the 0.1 level for several consecutive windows at the

beginning, with the improvement in the explanatory power of the model at a maximum 0.07 level; our interpretation is that the causality is nonlinear.

Figure 6. 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 (red line), when a model with one lag is considered. The right chart shows the scatter plot of *p*-values and the value of the causality measure.
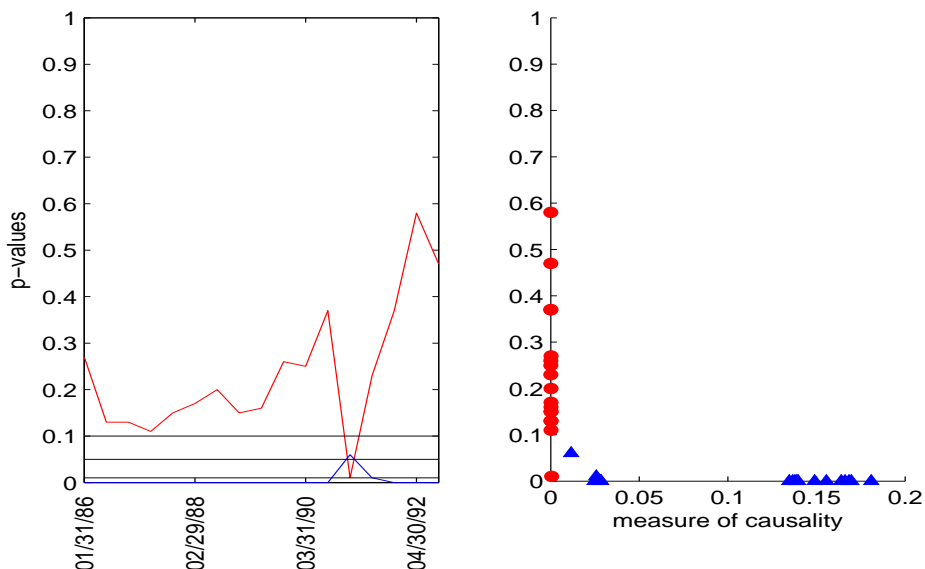

Figure 7. Linear Geweke's measure of causality. (**Left**) Sets of *p*-values for the hypothesis of statistical causality in the direction U.S. consumer price index → one-month Libor (blue line) or the other way round (red line), when a model with a linear kernel and Lag 1 is considered. (**Right**) Scatter plot of *p*-value and value of the causality measure.
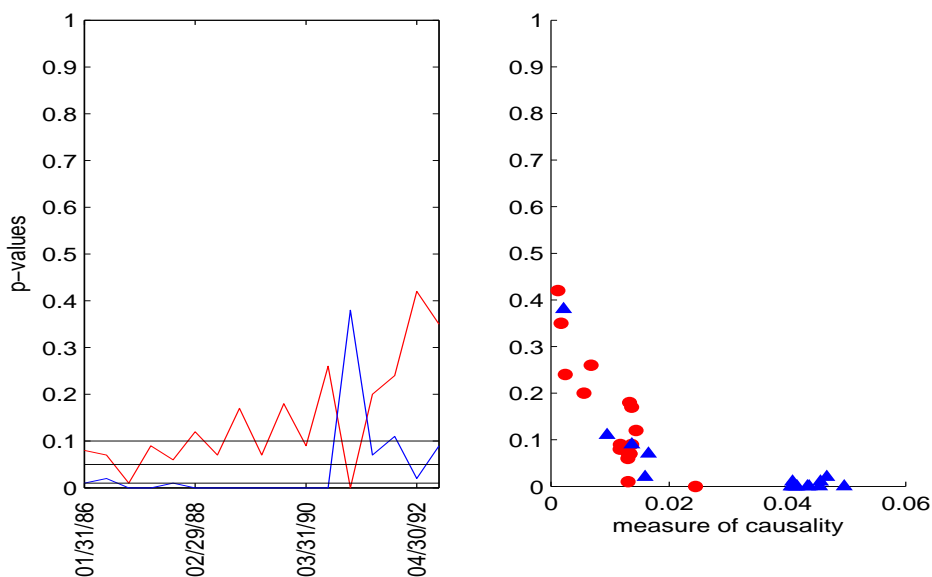
Figure 8. Kernelised Geweke's measure of causality. (**Left**) Sets of *p*-values for the hypothesis of statistical causality in the direction U.S. CPI → one-month Libor (blue line) or the other way round (red line), when the model with the Gaussian kernel and Lag 2 is considered; (**Right**) Scatter plot of the *p*-value and the value of the causality measure.
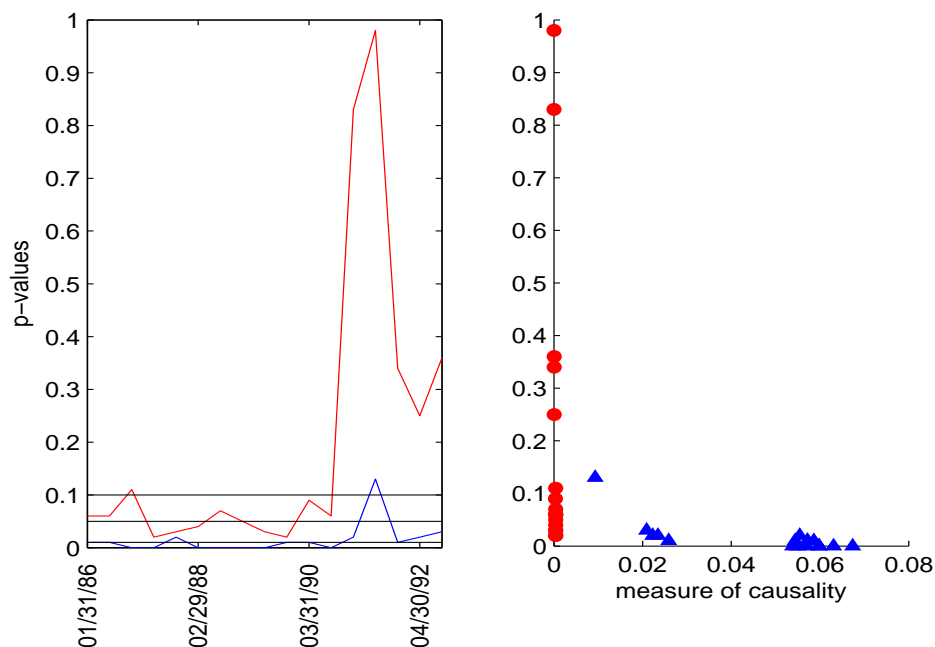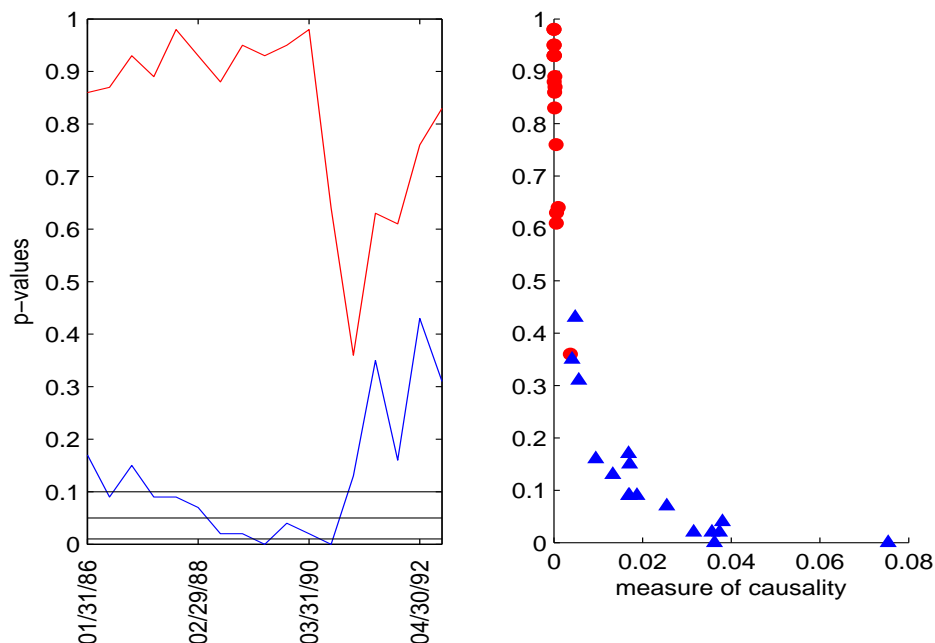

Figure 9. Linear Geweke's measure of causality. (**Left**) Sets of *p*-values for the hypothesis of statistical causality in the direction U.S. CPI → one-month Libor (blue line) or the other way round (red line), when model with a linear kernel and Lag 7 is considered; (**Right**) Scatter plot of the *p*-value and the value of the causality measure.

The results for the second lag, given in Figure 8, are no longer as clear as for Lag 1 in Figure 6 (Gaussian kernel 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 just below 0.08. Using a linear kernel, we obtain less clear results, and our interpretation is that the causal direction CPI $\to$ Libor is stronger, but there might be some feedback, as well.

Figure 9 presents the results of using a linear kernel, which shows 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 obvious reason why the linear kernel performed much better than the Gaussian kernel for these large lags. We offer the interpretation that no nonlinear causality was strong enough and consistent enough and that this was further obscured by using a nonlinear kernel. The conclusion here is that model selection is an important aspect of detecting causality and needs further research.

In our analysis, we did not obtain significant results for transfer entropy or HSNCIC. The results for Lag 1 for transfer entropy and HSNCIC are shown in Figures 10 and 11, respectively. For Lag 1, there was a significant statistical causality in the direction U.S. CPI $\to$ one-month Libor supported by both Geweke's measures. This is barely seen for transfer entropy and HSNCIC. *p*-values for transfer entropy are at a level that only slightly departs from a random effect, and for HSNCIC, they are often significant; however, the two directions are not well separated. The results for higher lags were often even more difficult to interpret. We must stress that the different implementation of transfer entropy and parameter choice for HSNCIC might result in better performance (please refer to Sections 5.1 and 5.2).

Figure 10. Transfer entropy. (**Left**) sets of *p*-values for the hypothesis of statistical causality in the direction U.S. CPI $\to$ one-month Libor (blue line) or the other way round (red line), when Lag 1 is considered; (**Right**) Scatter plot of the *p*-value and the value of the causality measure.
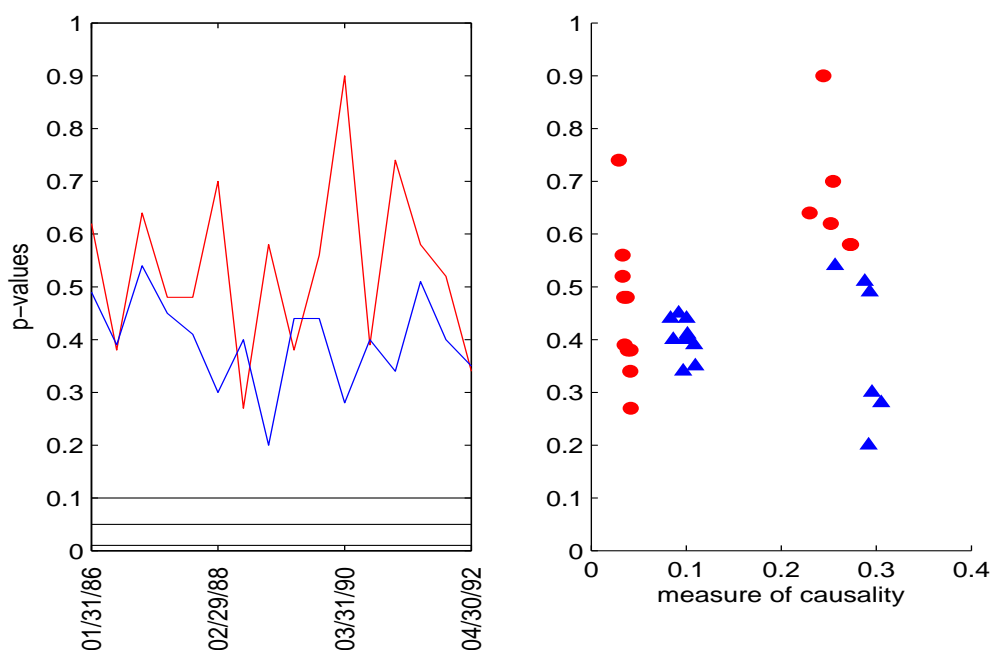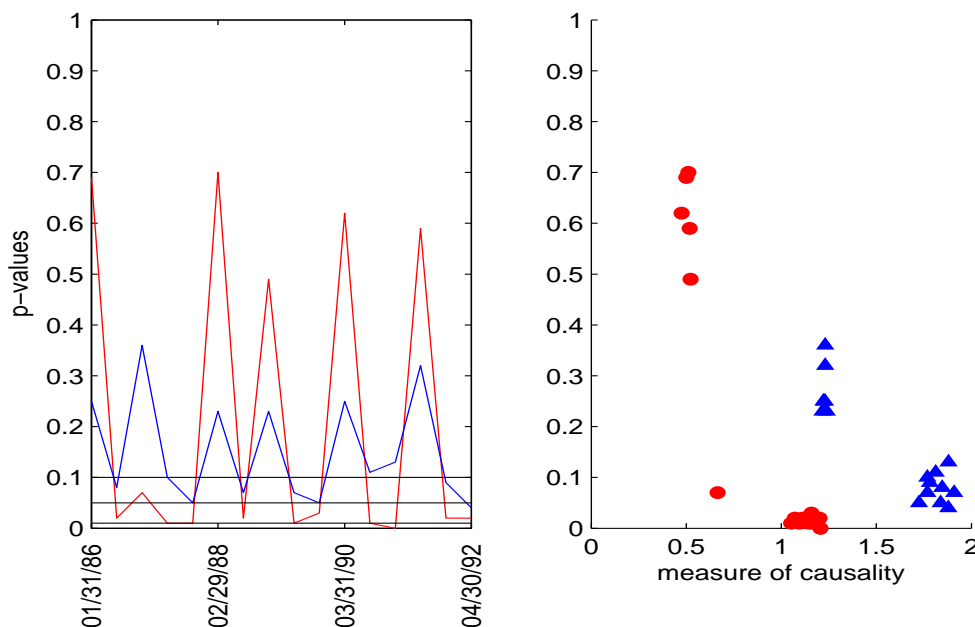
Figure 11. HSNCIC. (**Left**) sets of *p*-values for the hypothesis of statistical causality in the direction U.S. CPI $\rightarrow$ one-month Libor (blue line) or the other way round (red line), when Lag 1 is considered; (**Right**) Scatter plot of the *p*-value and the value of the causality measure.


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". Our expectation was that S&P should be leading. We used daily data for the period July 18, 2008–October 18, 2013, from Thomson Reuters. We studied the pairwise dependence between the currencies and S&P, and we also analysed the results of adding the Chicago Board Options Exchange Market Volatility Index (VIX) as side information. In all of the cases, we used logarithmic returns.

Figure 12 presents the results of applying kernelised Geweke's measure with a Gaussian kernel. The plots show series of *p*-values for a moving window of a length of 250 data points (days), moving each window by 25 points. Unlike in the previous case of interest rates and inflation, there is little actual difference between the linear and Gaussian kernel methods. However, in a few cases, employing a Gaussian kernel results in better separation of the two directions, especially CADCHF $\rightarrow$ S&P and S&P $\rightarrow$ CADCHF given VIX.

Excepting CADCHF, all currency pairs exhibit similar behaviour when analysed for the causal effect on the S&P. This behaviour consists of a small number of windows for which a causal relationship is significant at a *p*-value below 0.1, but that does not persist. CADCHF is the only currency with a 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 12. Sets of *p*-values for the hypothesis that an exchange rate causes the equity index, S&P (blue), or the other way round (red).
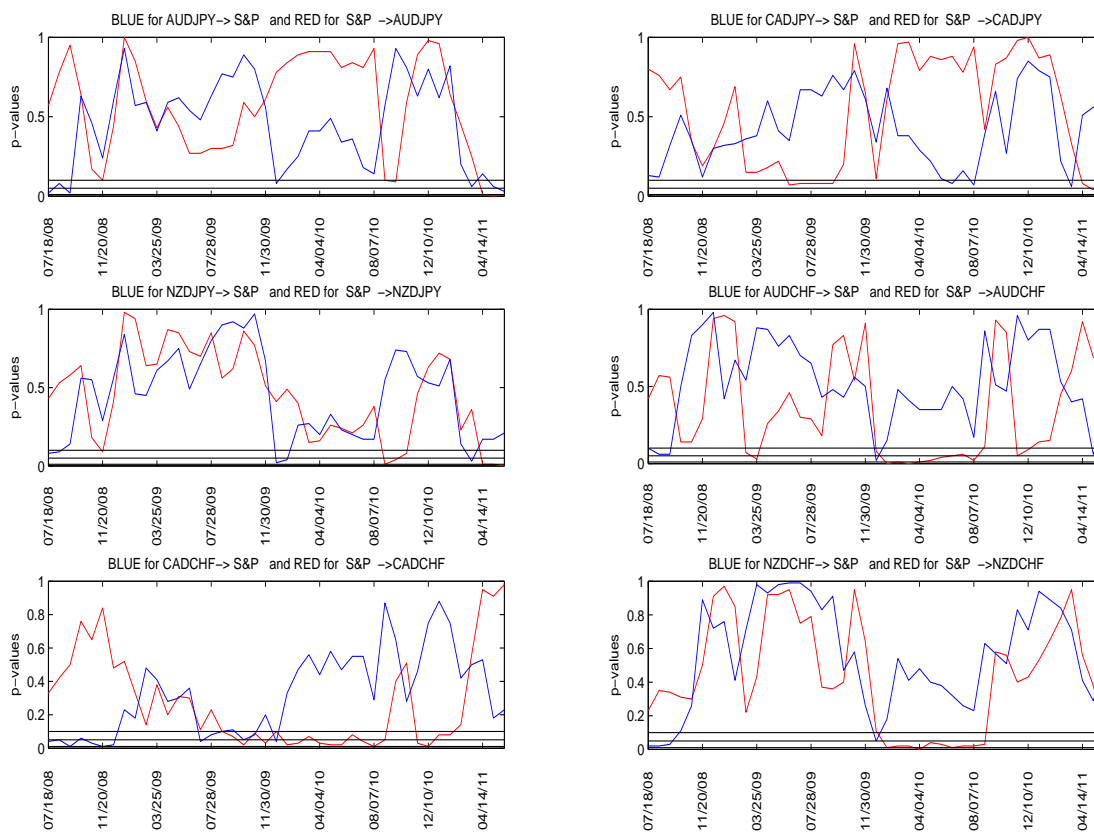

We obtained all of the main "regimes": Periods when either one of the exchange rates or S&P had more explanatory power (*p*-values for one direction were much lower than for the other) and periods when both exhibited low or both exhibited high *p*-values. *p*-values close to one did not necessarily mean purely a 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 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 the higher likelihood of interpreting correlation as causality. Furthermore, both of these biases can result from assuming a model with a lag different from that of the data. 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 these biases.

Figure 13 shows similar information as in Figure 12, but taking into consideration VIX as 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 do not show the disappearance of a causal effect after including VIX. While the patterns do not 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 13. Sets of *p*-values for the hypothesis that an exchange rate causes the equity index, S&P, given the Volatility Index (VIX) as side information (blue) or the other way round (red).
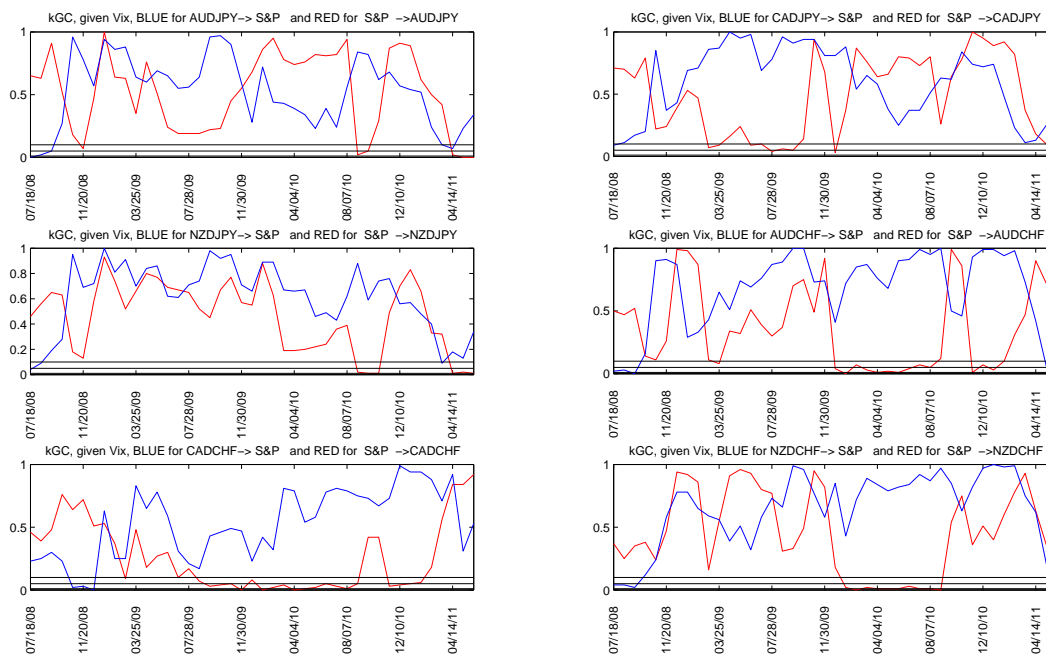
5. Discussion and Perspectives

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 [48]. A classical example from economics, known as "Milton Friedman's thermostat", describes how a lack of correlation is often confused with a lack of causation in the context of the evaluation of the Federal Reserve [49]. Secondly, often, questions are formulated in terms of (symmetrical) dependence, because it involves established methods and allows a clear interpretation. This could be a case in many risk management applications where the question of what causes losses should be central, but is not commonly addressed with causal methods [50]. The tools for quantifying causality that are currently being developed can help to better quantify causal inference and better understand the results.

In this section, we provide a critique of the methods to help understand their weaknesses and to enable the reader to choose the most appropriate method for the intended use. This will also set out the 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. Suggestions for future research directions conclude the section.

5.1. Theoretical Differences

Linearity *versus* nonlinearity: The original Granger causality and its Geweke's measure formulation were 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 or Gaussianity, linear methods should arguably not be used to analyse them. In practice, requirements on the size of the data sets 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 the reduction of dimensionality and repeated information. However, differentiation between direct and indirect causality is not necessarily well defined. This is because adding a conditioning variable can both introduce, as well as remove the dependence between variables [51]. 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 the defining of 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 [9], 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 are conditioned on side information and, therefore, are geared towards picking up only the direct cause; however, this does not work as intended, as we noticed that HSNCIC is extremely sensitive to the 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 [52,53], 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 inferred in relation 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 dependency. None of the methods we discuss in this paper is capable of managing several simultaneous types of dependency, be it instantaneous coupling, linear or nonlinear causality. We refer the interested reader to the literature on modelling Granger causality and transfer entropy in the frequency domain or using filters [31,54,55].

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 the kernelised Geweke's measure can somewhat mitigate this with parameter selection. Granger causality for bivariate data has good statistical tests for significance, while the others do not and need permutation tests that are computationally expensive. Furthermore, in the case of ridge regression, there is another layer of optimising parameters, which is

also computationally expensive. Calculating kernels is also relatively computationally expensive (unless the data is high-dimensional), but they are robust for small samples.

The HSNCIC is shown to have a good estimator, which, in the limit of infinite data, does not depend on the type of kernel. Transfer entropy, on the other hand, suffers from issues connected to estimating a distribution: problems with a small sample size and high dimensionality. Choosing the right estimator can help reduce the problem. A detailed overview of the possible methods of estimation of entropy can be found in [35]. Trentool, one of the more popular open access toolboxes for calculating transfer entropy, uses a nearest neighbour technique to estimate joint and marginal probabilities, that was first proposed by Kraskov *et al*. [27,39,56]. The nearest neighbour technique is data efficient, adaptive and has minimal bias [35]. The important aspect of this approach is that it depends on a correct choice of embedding parameter and, therefore, does not allow for analysing the information transfer for arbitrary lags. It also involves additional computational cost and might be slower for low dimensional data. We tested Trentool on several data sets and found that the demands on the size of the sample were higher than for the naive histogram and the calculations were slower, with comparable results. The naive histogram, however, does not have good performance for higher dimensions [35], in which case, the nearest neighbour approach would be advised.

Non-stationarity: This is one of the most important areas for future research. All of the described measures suffer to some degree from an inability to deal with non-stationary data. Granger causality in the original, linear formulation, is the only measure that explicitly assumes stationarity (more precisely, covariance stationarity [5,7]), and the asymptotic theory is developed for that case. Geweke describes in [57] special cases of non-stationary processes that can still be analysed within the standard framework and corresponding literature on adapting the linear Granger causality framework to the case of integrated or cointegrated processes [58]. In all of those cases, the type of non-stationarity needs to be known, and that is a potential source of new biases [58]. The GCCA toolbox [59] 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 non-stationarity: the ADF test (Augmented Dickey–Fuller) and the KPSS test (Kwiatkowski, Phillips, Schmidt, Shin). For managing non-stationarity, the GCCA toolbox manual [31] 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.

The other measures described in this article do not explicitly assume stationarity; however, some assumptions about stationarity are necessary for the methods to work correctly. Schreiber developed transfer entropy under the assumption that an analysed system can be approximated by stationary Markov processes [26]. Transfer entropy in practice can be affected if the time series is highly non-stationary, as the reliability of the estimation of probability densities will be biased [39], but non-stationarity, due to the slow change of the parameters, does not have to be a problem [60]. For the other two methods, the kernelised Geweke's measure and HSNCIC, the results for estimator convergence are available only for stationary data, according to our knowledge. However, the asymptotic results for HSNCIC have been developed for the too restrictive case of i.i.d. data [61]. The results for kernel ridge regression given by [62] have been developed for alpha-mixing data.

Choice of parameters: Each of the methods requires parameter selection; an issue related to model selection described in Section 5.2. All of the methods need a choice of the number of lags (lag order), while kernel methods additionally require the choice of kernel, kernel parameter (kernel size) and regularisation parameter.

In the case of the Gaussian kernel, the effect of the kernel size on the smoothing of the data can be understood as follows [63,64]. 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. If we express a kernel as Taylor series expansions, using the basis $1, u, u^2, u^3, ...$, the random variables, X and Y, can be expressed in RKHS by:

$$\Phi(X) = k(X, \cdot) \sim (1, c_1X, c_2X^2, c_3X^3, ...)^T$$
$$\Phi(Y) = k(Y, \cdot) \sim (1, c_1Y, c_2Y^2, c_3Y^3, ...)^T \tag{28}$$

therefore, the kernel function can be expressed as follows:

$$k(x, y) = 1 + c_1xy + c_2x^2y^2 + c_3x^3y^3 + ... \tag{29}$$

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

$$\Sigma_{XY} \sim \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & c_1^2Cov[X, Y] & c_1c_2Cov[X, Y^2] & c_1c_3Cov[X, Y^3] & ... \\ 0 & c_2c_1Cov[X^2, Y] & c_2^2Cov[X^2, Y^2] & c_2c_3Cov[X^2, Y^3] & ... \\ 0 & c_3c_1Cov[X^3, Y] & c_3c_2Cov[X^3, Y^2] & c_3^2Cov[X^3, Y^3] & ... \\ 0 & ... & ... & ... & ... \end{pmatrix} \tag{30}$$

According to Fukumizu *et al.* [15], 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. The 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 and the binning size in the case of the naive histogram approach.

Another important aspect is the choice of lag order and the number of lags. We observed in Section 3.2.1. that the two Geweke's measures were not sensitive to the choice of lags, and we were able to correctly recognise causality both in the case of the smaller and bigger lag ranges used. The two other measures, however, behaved differently. HSNCIC is often not able to observe causality in the case of more lags analysed at a time, but performed well for single lags. Transfer entropy flagged spurious causality in one case where the lag was far from the "true" one. However, for real data, with a more complex structure, the choice of lag is likely to be important for all measures (see Section 5.2).

5.2. Model Selection

For the kernel measures, we observed that model selection was an important issue. In general, the choice of kernel influences the smoothness of the class of functions considered, while the choice of regulariser controls the trade-off between the smoothness of the function and the error of the fit.

Underfitting can be a consequence of a too large regulariser and a too large kernel size (in the case of a Gaussian kernel); conversely, overfitting can be a consequence of a too small regulariser and a too small kernel size. One of the methods suggested to help with model selection is cross-validation [9]. This method is particularly popular and convenient for the selection of kernel size and regulariser in the ridge regression (see Appendix E). Given non-stationary data, it would seem reasonable to fit the parameters; however, we concluded that cross-validation was too expensive in the computational sense and did not provide the expected benefits.

Another aspect of model selection (and the choice of parameters) is the determination of an appropriate lag order. For kernel methods, increasing the number of lags does not increase the dimensionality of the problem, as could be expected in the case of the methods representing the data explicitly. As described in Section 2.1.3., in the case of the kernelised Geweke's measure, increasing the number of lags decreases the dimensionality of the problem, due to the fact that the data is represented in terms of $(n-p) \times (n-p)$ pairwise comparisons, where n is the number of observations and p the number of lags. On the other hand, increasing the number of lags will decrease the number of degrees of freedom. This decrease will be less pronounced for kernel methods which allocate smaller weights to higher lags (as is the case in the Gaussian kernel, but not for the linear kernel). Apart from cross-validation, the other approaches to choosing the lag order suggested in the literature are based on the analysis of the autocorrelation function or partial autocorrelation [27,65].

We feel that more research is needed on 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 was described in Section 3.1 involves the destruction of all types of dependency, not just causal dependence. In practice, this means that, for example, the existence of instantaneous coupling can result in the incorrect deduction of causal inference, if the improvement in prediction due to the existence of causality is confused with improvement, due to instantaneous coupling. Nevertheless, simplicity is the deciding factor in favour of permutation tests over other approaches.

Several authors [9,18,25] 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 in 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 is big and the assumption of stationarity is reasonable, but that was not the case for our data. We instead decided to report *p*-values for an overlapping running window. This allows us to additionally assess the consistency of the results and does not require us to choose the same significance rate for all of the windows.

5.4. Perspectives

In the discussion, we highlighted many areas that still require more research. The kernelised Geweke's measure, transfer entropy and HSNCIC detect nonlinear dependence better than the original Granger

causality, but do not improve on its other weakness: non-stationarity. Ridge regression is a convenient tool in online learning, and it could prove helpful in dealing with non-stationarity [9]. This is clearly an area worth exploring.

Table 3. The summary of the 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 the measure with and without side information
kernelised Geweke's	to some extent by comparing the measure with and without side information
transfer entropy	not able to (consider partial transfer entropy)
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 (Augmented Dickey–Fuller), KPSS (Kwiatkowski, Phillips, Schmidt, Shin) 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

Crucially for applications to financial data, more needs to be understood about measuring causality in time series with several types of dependency. We are not aware of any study that addresses 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 frequency-based decomposition. A different type of filtering is decomposition into negative and positive shocks, for example Hatemi-J proposed an "asymmetric causality measure" based on Granger causality [66].

The third main direction of suggested research is building causal networks. There is a substantial body of literature about causal networks for intervention-based causality, 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 [29,46,67,68]. 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 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 conclude that no measure is clearly superior to the others. We believe, however, that the 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 real data sets, we were able to identify causal directions that demonstrated some consistency between methods and time windows and that did not conflict with the economic rationale. The two experiments identified a range of limitations that need to be addressed to allow for a wider application of any the methods to financial data. Furthermore, neither of the data sets contained higher frequency data, and working with a high frequency is likely to produce additional complications.

A separate question that we only briefly touched upon is the relevance and practicality of using any causality measure. This is a question lying largely in the domain of the philosophy of science. Ultimately, it is the interpretation of researchers and their confidence in the data that makes it possible to label a relationship as causal rather than only statistically causal. However, 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 since Granger was awarded the Nobel prize in 2003. There is little literature on using nonlinear generalisations of the Granger causality in finance or economics. We believe that it has great potential, on the one hand, and still many questions to be answered, on the other. While we expect that some of the problems could be addressed with an online learning approach and data filtering, more research on dealing with non-stationarity, noisy data and optimal parameter selection is required.

Appendices

A. Solving Ridge Regression

The regularised cost function is Equation (9):

$$\beta^* = \underset{\beta}{\text{argmin}} \frac{1}{m} \sum_{i=p+1}^{n} ((w_{i-1}^{i-p})^T \beta - x_i)^2 + \gamma \beta^T \beta \tag{31}$$

Now solving Equation (9) gives:

$$\begin{aligned}
\mathcal{L} &= \frac{1}{m} \sum_{i=p+1}^{n} ((w_{i-1}^{i-p})^T \beta - x_i)^2 + \gamma \beta^T \beta \\
&= \frac{1}{m} (\mathbf{W}\beta - x)^T (\mathbf{W}\beta - x) + \gamma \beta^T \beta = \\
&= \frac{1}{m} (\beta^T \mathbf{W}^T \mathbf{W}\beta - 2x^T \mathbf{W}\beta + x^T x) + \gamma \beta^T \beta \\
\frac{\partial \mathcal{L}}{\partial \beta} &= \frac{1}{m} (2\mathbf{W}^T \mathbf{W}\beta - \mathbf{W}^T x) + 2\gamma \beta = 0 \Leftrightarrow \\
&\Leftrightarrow \mathbf{W}^T \mathbf{W}\beta^* + \gamma m \beta^* = \mathbf{W}^T x \Leftrightarrow \\
&\Leftrightarrow \beta^* = (\mathbf{W}^T \mathbf{W} + \gamma m I_m)^{-1} \mathbf{W}^T x
\end{aligned} \tag{32}$$

where I_m is an $m \times m$ identity matrix.

The weights, β^*, are called the primal solution, and the next step is to introduce the dual solution weights.

$$\mathbf{W}^T \mathbf{W}\beta^* + \gamma m \beta^* = \mathbf{W}^T x \Leftrightarrow \beta^* = \frac{1}{\gamma m} \mathbf{W}^T (x - \mathbf{W}\beta^*) \tag{33}$$

so for some $\alpha^* \in \mathbb{R}^n$, we can write that:

$$\beta^* = \mathbf{W}^T \alpha^* \tag{34}$$

From the two sets of equation above, we get that:

$$\begin{aligned}
\alpha^* &= \frac{1}{\gamma m} (x - \mathbf{W}\beta^*) \\
&\Leftrightarrow \gamma m \alpha^* = x - \mathbf{W}\beta^* = x - \mathbf{W}\mathbf{W}^T \alpha^* \Leftrightarrow \\
&\Leftrightarrow (\mathbf{W}\mathbf{W}^T + \gamma m I_m)\alpha^* = x
\end{aligned} \tag{35}$$

This gives the desired form for the dual weights:

$$\alpha^* = (\mathbf{W}\mathbf{W}^T + \gamma m I_m)^{-1} x \tag{36}$$

which depend on the regularisation parameter, γ.

B. Background from Functional Analysis and Hilbert Spaces

The definitions and theorems below follow [13,18,69]. All vector spaces will be over $\mathbb{R}$, rather than $\mathbb{C}$; however, 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 $\langle \cdot, \cdot \rangle_{\mathcal{F}} : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ is said to be an inner product on $\mathcal{F}$ if:*

$$\begin{aligned}
&\textit{(i)} \quad \langle f_1 + f_2, f \rangle = \langle f_1, f \rangle + \langle f_2, f \rangle, \quad \textit{for all } f, f_1, f_2 \in \mathcal{F} \\
&\textit{(ii)} \quad \langle \alpha f_1, f_2 \rangle = \alpha \langle f_1, f_2 \rangle \quad \textit{for all } f_1, f_2 \in \mathcal{F}, \alpha \in \mathbb{R} \\
&\textit{(iii)} \quad \langle f_1, f_2 \rangle = \langle f_2, f_1 \rangle \quad \textit{for all } f_1, f_2 \in \mathcal{F} \\
&\textit{(iv)} \quad \langle f, f \rangle \geq 0 \quad \textit{and } \langle f, f \rangle = 0 \quad \textit{if and only if } f = 0
\end{aligned} \tag{37}$$

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 the metric induced by the inner product is complete [70].*

One of the fundamental concepts of functional analysis that we will utilise is that of a continuous linear operator: for two vector spaces, $\mathcal{F}$ and $\mathcal{G}$, over $\mathbb{R}$, a map $T : \mathcal{F} \to \mathcal{G}$ 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 the standard notational convention $Tf := T(f)$.

The following three conditions can be proven to be equivalent: (1) linear operator T is continuous; (2) T is continuous at zero; and (3) T is bounded [71]. This result, along with the Riesz representation theorem given later, is fundamental for the theory of reproducing kernel Hilbert spaces. It should be emphasised that while the operators we use, such as the 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 operators $T : \mathcal{F} \to \mathbb{R}$.

Theorem 2 *(Riesz representation theorem) In a Hilbert space, $\mathcal{F}$, all continuous linear functionals [72] are of the form $\langle \cdot, f \rangle$, for some $f \in \mathcal{F}$.*

In Appendix A, we used the "kernel trick" without explaining why it was permissible. The explanation is given below as the representer theorem. The theorem refers 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. The risk, $\mathcal{R}_{L,S}$, associated with the loss, L, and data sample S is defined as the average future loss of the prediction function, f.

Theorem 3 *(Representer theorem) [69] Let $L : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ 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:*

$$\lambda \|f_{S,\lambda}\|_{\mathcal{F}}^2 + \mathcal{R}_{L,S}(f_{S,\lambda}) = \inf_{f \in \mathcal{F}} \lambda \|f\|_{\mathcal{F}}^2 + \mathcal{R}_{L,S}(f) \tag{38}$$

In addition, there exist $\alpha_1, \cdots \alpha_n \in \mathbb{R}$, such that:

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

Below, we present definitions that are the 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 $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let $\mathcal{G}$ be an RKHS of functions from $\mathcal{Y}$ to $\mathbb{R}$, induced by strictly positive kernel $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ [73]. Denote by $C : \mathcal{G} \to \mathcal{F}$ 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 [13,18], let $\mathcal{H}_\mathcal{W}$ denote the RKHS induced by a strictly positive kernel $k_\mathcal{W} : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$. Let X be a random variable on $\mathcal{X}$, Y be a random variable on $\mathcal{Y}$ and (X, Y) be 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 relevant $\sigma-$fields. The marginal distributions are denoted by P_X, P_Y and the joint distribution of (X, Y) by P_{XY}. The expectations, $\mathbf{E}_X$, $\mathbf{E}_Y$ and $\mathbf{E}_{XY}$, denote the expectations over P_X, P_Y and P_{XY}, respectively. To ensure $\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{Y}$ are included in, respectively, $L^2(P_X)$ and $L^2(P_Y)$, we consider only random vectors (X, Y), such that the expectations, $\mathbf{E}_X[k_\mathcal{X}(X, X)]$ and $\mathbf{E}_Y[k_\mathcal{Y}(Y, Y)]$, are finite.

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

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

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

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 $f \otimes g : \mathcal{G} \to \mathcal{F}$ 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 the application of an operator: $(f \otimes g)h$ instead of $(f \otimes g)(h)$. The second one relates to multiplication by a scalar, and we write $f\langle g, h \rangle_\mathcal{G}$ instead of $f \cdot \langle g, h \rangle_\mathcal{G}$.

The Hilbert–Schmidt norm of the tensor product can be calculated as:

$$\begin{aligned} \|f \otimes g\|_{HS}^2 = \langle f \otimes g, f \otimes g \rangle_{HS} = \langle f, (f \otimes g)g \rangle_\mathcal{F} \\ = \langle f, f \rangle_\mathcal{F} \langle g, g \rangle_\mathcal{G} = \|f\|_\mathcal{F}^2 \|g\|_\mathcal{G}^2 \end{aligned} \tag{43}$$

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

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

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

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

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

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

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

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

C. Hilbert–Schmidt Independence Criterion (HSIC)

As in Section 2.1.4., following [13,18], let $\mathcal{F}_{\mathcal{X}}, \mathcal{F}_{\mathcal{Y}}$ denote the RKHS induced by strictly positive kernels $k_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $k_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Let X be a random variable on $\mathcal{X}$, Y be a random variable on $\mathcal{Y}$ and (X,Y) be a 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 $\mathcal{F}_{\mathcal{X}}, \mathcal{F}_{\mathcal{Y}}, P_X, P_Y$ as introduced earlier, we define the Hilbert–Schmidt independence criterion as the squared Hilbert–Schmidt norm of the cross-covariance operator, Σ_{XY}:*

$$HSIC(P_{XY}, \mathcal{F}_{\mathcal{X}}, \mathcal{F}_{\mathcal{Y}}) := \|\Sigma_{XY}\|_{HS}^2 \tag{47}$$

We cite without proof the following lemma from [13]:

Lemma 1 *(HSIC in kernel notation)*

$$
\begin{aligned}
HSIC(P_{XY}, \mathcal{F}_{\mathcal{X}}, \mathcal{F}_{\mathcal{Y}}) := & \boldsymbol{E}_{X,X',Y,Y'}[k_{\mathcal{X}}(X,X')k_{\mathcal{Y}}(Y,Y')] + \boldsymbol{E}_{X,X'}[k_{\mathcal{X}}(X,X')]\boldsymbol{E}_{Y,Y'}[k_{\mathcal{Y}}(Y,Y')] \\
& -2\boldsymbol{E}_{X,Y}[\boldsymbol{E}_{X'}[k_{\mathcal{X}}(X,X')]\boldsymbol{E}_{Y'}[k_{\mathcal{Y}}(Y,Y')]]
\end{aligned}
\tag{48}
$$

where X, X' and Y, Y' are independent copies of the same random variable.

D. Estimator of HSNCIC

Empirical mean elements:

$$
\begin{aligned}
\hat{m}_X^{(n)} &= \frac{1}{n} \sum_{i=1}^n k_{\mathcal{X}}(\cdot, X_i) \\
\hat{m}_Y^{(n)} &= \frac{1}{n} \sum_{i=1}^n k_{\mathcal{Y}}(\cdot, Y_i)
\end{aligned}
\tag{49}
$$

Empirical cross-covariance operator:

$$
\begin{aligned}
\hat{\Sigma}_{XY}^{(n)} &= \frac{1}{n} \sum_{i=1}^n (k_{\mathcal{Y}}(\cdot, Y_i) - \hat{m}_Y^{(n)}) \langle k_{\mathcal{X}}(\cdot, X_i) - \hat{m}_X^{(n)}, \cdot \rangle_{\mathcal{H}_{\mathcal{X}}} \\
&= \frac{1}{n} \sum_{i=1}^n \{k_{\mathcal{Y}}(\cdot, Y_i) - \hat{m}_Y^{(n)}\} \otimes \{k_{\mathcal{X}}(\cdot, X_i) - \hat{m}_X^{(n)}\}
\end{aligned}
\tag{50}
$$

Empirical normalised cross-covariance operator:

$$\hat{V}_{XY}^{(n)} = (\hat{\Sigma}_{XX}^{(n)} + n\lambda I_n)^{-1/2} \hat{\Sigma}_{XY}^{(n)} (\hat{\Sigma}_{YY}^{(n)} + n\lambda I_n)^{-1/2} \tag{51}$$

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

Empirical normalised conditional cross-covariance operator:

$$\hat{V}_{XY|Z}^{(n)} = \hat{V}_{XY}^{(n)} - \hat{V}_{XZ}^{(n)} \hat{V}_{ZY}^{(n)} \tag{52}$$

For U symbolising any of the variables, (XZ), (YZ) or Z, we denote by K_U a centred Gram matrix, such that each elements equal to: $K_{U,ij} = \langle k_\mathcal{U}(\cdot, U_i) - \hat{m}_U^{(n)}, k_\mathcal{U}(\cdot, U_j) - \hat{m}_U^{(n)} \rangle_{\mathcal{H}_\mathcal{U}}$; let $R_U = K_U(K_U + n\lambda I)^{-1}$. With this notation, the empirical estimation of HSNCIC can be written as:

$$HSNCIC_n := Tr[R_{(XZ)}R_{(YZ)} - 2R_{(XZ)}R_{(YZ)}R_Z + R_{(XZ)}R_Z R_{(YZ)}R_Z] \tag{53}$$

E. Cross-Validation Procedure

Obtaining a kernel (or more precisely, a Gram matrix) is computationally intensive. Performing cross-validation requires the calculation of two 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 the validation points are subsequent), calculating the kernel for the whole (but appropriately ordered) dataset and selecting the appropriate part of the kernel for testing and validation:

$$\bar{K} = \begin{array}{|c|c|} \hline K(W_{train}, W_{train}) & K(W_{train}, W_{val}) \\ \hline K(W_{val}, W_{train}) & K(W_{val}, W_{val}) \\ \hline \end{array} \tag{54}$$

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

$$\bar{K}_{val} = K(W_{val}, W_{train}) \tag{55}$$

Such an 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:

$$\frac{1}{m}(\bar{K}_{val}\alpha^* - x)^T(\bar{K}_{val}\alpha^* - x) \tag{56}$$

Even with an effective way of calculating kernels, cross-validation is still expensive. As described below, to obtain a 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 practice, in order to run several experiments, using a number of measures in a reasonable amount of time, a reasonable compromise is not to perform the cross-validation after every step, but once per experiment and use those parameters in all trials.

We believe that one of the strengths of the kernelised Geweke's measure, and one of the reasons why kernels are often used for online learning, lies in the fact that it is possible to optimise the parameters, but the parameters do not have to be optimised each time.

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 the Gaussian kernel, which is defined as follows:

$$k(x, y) = exp(-\frac{\|x - y\|^2}{\sigma^2}) \tag{57}$$

and the linear kernel is defined as $k(x, y) = x^T y$.

We use a randomized five-fold cross-validation to choose the optimal parameter, γ, for the regularisation and the kernel parameters. Let $(x_t, y_t, z_t), t = 1, ..., n$ be the time series. We want to

calculate $G_{y \to x \| z}$. Based on the given time series, we create a learning set with the lag (embedding) equal to p, and we prepare a learning set following the notation from Section 2.1.3. : (x_i, w_{i-1}^{i-p}), for $i = p + 1, ..., n$. The learning set is split randomly into five subsets of equal size. For each $k = 1, ..., 5$, we obtain a k-th validation set and a k-th testing set that contains all data points that do not belong to the k-th validation set.

Next, a grid is created, given a range of values for the parameter, γ, 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 the grid with the 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, an unnecessarily big grid is undesirable.

Acknowledgements

Author Contributions

All authors contributed to the conception and design of the study, the collection and analysis of the data and the discussion of the results.

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. *J. Econ. Dyn. Control* **1980**, *2*, 329–352.
3. Norbert, W. The Theory of Prediction. In *Modern Mathematics for Engineers*; Beckenbach, E.F., Ed.; McGraw-Hill: New York, NY, USA, 1956; Volume 1.

4. Granger, C. Economic processes involving feedback. *Inform. Contr.* **1963**, *6*, 28–48.

5. Granger, C.W.J. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* **1969**, *37*, 424–438.

6. Geweke, J. Measurement of Linear Dependence and Feedback between Multiple Time Series. *J. Am. Stat. Assoc.* **1982**, *77*, 304–313.

7. Geweke, J.F. Measures of Conditional Linear Dependence and Feedback Between Time Series. *J. Am. Stat. Assoc.* **1984**, *79*, 907–915.

8. A 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 here "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

9. Amblard, P.; Vincent, R.; Michel, O.; Richard, C. Kernelizing Geweke's Measures of Granger Causality. In Proceedings of the 2012 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), Santander, Spain, September 2012; pp. 1–6.

10. Amblard, P.O.; Michel, O.J.J. The Relation between Granger Causality and Directed Information Theory: A Review. *Entropy* **2012**, *15*, 113–143.

11. However, such confusion might not necessarily 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 either 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.

12. Gretton, A.; Herbrich, R.; Smola, A.; Bousquet, O.; Schoelkopf, B. Kernel Methods for Measuring Independence. *J. Mach. Learn. Res.* **2005**, *6*, 2075–2129.

13. Gretton, A.; Bousquet, O.; Smola, A.; Schoelkopf, B. Measuring Statistical Dependence with Hilbert–Schmidt Norms. In *Proceedings Algorithmic Learning Theory*; Springer-Verlag: Berlin/Heidelberg, Germany, 2005; p. 63–77.

14. Sun, X.; Janzing, D.; Schoelkopf, B.; Fukumizu, K. A Kernel-Based Causal Learning Algorithm. In Proceedings of the 24th International Conference on Machine Learning, Oregon, OR, USA, 21–23 June 2007; ACM: New York, NY, USA, 2007; pp. 855–862.

15. Fukumizu, K.; Gretton, A.; Sun, X.; Schoelkopf, B. Kernel Measures of Conditional Dependence. *Adv. NIPS*, **2007**, *20*, 489–496.

16. Gretton, A.; Fukumizu, K.; Teo, C.H.; Song, L.; Schoelkopf, B.; Smola, A.J. A kernel statistical test of independence; In *Advances in Neural Information Processing Systems 20: 21st Annual Conference on Neural Information Processing Systems 2007*; MIT Press: Cambridge, MA, USA, 2008.

17. Guyon, I.; Janzing, D.; Schoelkopf, B. Causality: Objectives and Assessment. *J. Mach. Learn. Res.* **2010**, *6* 1–42.

18. 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, Germany, 2008; pp. 440–455.

19. Marinazzo, D.; Pellicoro, M.; Stramaglia, S. Kernel-Granger causality and the analysis of dynamical networks. *Phys. Rev. E, Stat. Nonlinear Soft Matter Phys.* **2008**, *77*, 056215.

20. Schoelkopf, B.; Tsuda, K.; Vert, J.P. Kernel Methods in Computational Biology. MIT Press: Cambridge, MA, USA, 2004.

21. Hastie, T.; Tibshirani, R.; Friedman, J.H. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*; Springer: New York, NY, USA, 2009.

22. Renyi, A. On measures of dependence. *Acta Math. Acad. Sci. Hung.* **1959**, *10*, 441–451.

23. 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. In Proceedings of the Society for Artificial Intelligence and Statistics, Hastings, Barbados, January 2005.

24. Hofmann, T.; Schoelkopf, B.; Smola, A.J. Kernel methods in machine learning. *Ann. Stat.* **2008**, *36*, 1171–1220.

25. Seth, S.; Principe, J. Assessing Granger Non-Causality Using Nonparametric Measure of Conditional Independence. *IEEE Trans. Neural Netw. Learn. Syst.* **2012**, *23*, 47–59.

26. Schreiber, T. Measuring Information Transfer. *Phys. Rev. Lett.* **2000**, *85*, 461–464.

27. 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 Neurosci.* **2011**, *12*, 119.

28. Lungarella, M.; Ishiguro, K.; Kuniyoshi, Y.; Otsu, N. Methods for Quantifying the Causal Structure of Bivariate Time Series. *Int. J. Bifurc. Chaos* **2007**, *17*, 903–921.

29. Amblard, P.O.; Michel, O.J.J. On directed information theory and Granger causality graphs. *J. Comput. Neurosci.* **2011**, *30*, 7–16.

30. Barnett, L.; Barrett, A.B.; Seth, A.K. Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables. *Phys. Rev. Lett.* **2009**, *103*, 238701.

31. Seth, A.K. A MATLAB toolbox for Granger causal connectivity analysis. *J. Neurosci. Methods* **2010**, *186*, 262–273.

32. Anil Seth Code. Aviable online: http://www.sussex.ac.uk/Users/anils/aks_code.htm (accessed on 20 April 2014).

33. The Code is available at http://www.sohanseth.com/Home/publication/causmci (accessed on 20 April 2014).

34. The use of permutation, which is more general than rotation, is helpful when data is short or the analysed time windows are short.

35. Hlavackova-Schindler, K.; Palus, M.; Vejmelka, M.; Bhattacharya, J. Causality detection based on information-theoretic approaches in time series analysis. *Phys. Rep.* **2007**, *441*, 1–46.

36. Chávez, M.; Martinerie, J.; le van Quyen, M. Statistical assessment of nonlinear causality: Application to epileptic EEG signals. *J. Neurosci. Methods* **2003**, *124*, 113–128.

37. 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. In Proceedings of the Earth-Sun System Technology Conference (ESTC 2005), Orange County, CA, USA, January 2005.

38. Gourévitch, B.; Eggermont, J.J. Evaluating information transfer between auditory cortical neurons. *J. Neurophysiol.* **2007**, *97*, 2533–2543.

39. Vicente, R.; Wibral, M.; Lindner, M.; Pipa, G. Transfer entropy–a model-free measure of effective connectivity for the neurosciences. *J. Comput. Neurosci.* **2011**, *30*, 45–67.

40. Bressler, S.L.; Seth, A.K. Wiener-Granger causality: A well established methodology. *NeuroImage* **2011**, *58*, 323–329.

41. Alexander, C.; Wyeth, J. Cointegration and market integration: An application to the Indonesian rice market. *J. Dev. Stud.* **1994**, *30*, 303–334.

42. Cont, R. Long Range Dependence in Financial Markets. In *Fractals in Engineering*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 159–180.

43. Patton, A.J. Copula-Based Models for Financial Time Series. In *Handbook of Financial Time Series*; Mikosch, T., Kreiß, J.P., Davis, R.A., Andersen, T.G., Eds.; Springer: Berlin/Heidelberg, Germany, 2009; pp. 767–785.

44. 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., Mesiar, R., Calvo, T., Eds.; Number 228 in Advances in Intelligent Systems and Computing; Springer: Berlin/Heidelberg, Germany, 2013; pp. 17–22.

45. Barber, D. *Bayesian Reasoning and Machine Learning*; Cambridge University Press: Cambridge, UK, 2012.

46. Eichler, M. Granger causality and path diagrams for multivariate time series. *J. Econ.* **2007**, *137*, 334–353.

47. We use Granger's "strength of causality" for this purpose [6]. If we denote by Σ_1 and Σ_2 the variance of errors in Models 1 and 2, respectively, the strength of causality will be equal to $1 - |\Sigma_2|/|\Sigma_1|$. The Geweke's measure we use is given by $ln(|\Sigma_1|/|\Sigma_2|)$, which can be transformed to give the value of the strength of causality.

48. An example that had considerable media coverage in many countries is the paper from 2012 by scientists from the Harvard School of Public Health about the effect of the consumption of red meat on mortality [74].

49. Friedman, M. The Fed's Thermostat. Available online: http://online.wsj.com/news/articles/SB106125694925954100 (accessed on 20 April 2014).

50. Fenton, N.; Neil, M. *Risk Assessment and Decision Analysis with Bayesian Networks*; CRC Press: Boca Raton, FL, USA, 2012.

51. Hsiao, C. Autoregressive modelling and causal ordering of economic variables. *J. Econ. Dyn. Control* **1982**, *4*, 243–259.

52. Papana, A.; Kyrtsou, C.; Kugiumtzis, D.; Diks, C. Simulation Study of Direct Causality Measures in Multivariate Time Series. *Entropy* **2013**, *15*, 2635–2661.

53. Kugiumtzis, D. Partial transfer entropy on rank vectors. *Eur. Phys. J. Special Top.* **2013**, *222*, 401–420.

54. Lungarella, M.; Pitti, A.; Kuniyoshi, Y. Information transfer at multiple scales. *Phys. Rev. E, Stat. Nonlinear Soft Matter Phys.* **2007**, *76*, 056117.

55. Dhamala, M.; Rangarajan, G.; Ding, M. Estimating Granger causality from fourier and wavelet transforms of time series data. *Phys. Rev. Lett.* **2008**, *100*, 018701.

56. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E, Stat. Nonlinear Soft Matter Phys.* **2004**, *69*, 066138.

57. Geweke, J. Chapter 19 Inference and Causality in Economic Time Series Models. In *Handbook of Econometrics*; Griliches, Z., Intriligator, M.D., Eds.; Elsevier: Amsterdam, The Netherlands, 1984; Volume 2, pp. 1101–1144.

58. Toda, H.Y.; Yamamoto, T. Statistical inference in vector autoregressions with possibly integrated processes. *J. Econ.* **1995**, *66*, 225–250.

59. Code can be requested at: http://www.sussex.ac.uk/Users/anils/aks_code.htm.

60. Gomez-Herrero, G.; Wu, W.; Rutanen, K.; Soriano, M.C.; Pipa, G.; Vicente, R. Assessing coupling dynamics from an ensemble of time series. **2010**, *rXiv:1008.0539* .

61. We believe that the generalisation from i.i.d. data to alpha-mixing can be done similarly as for the HSIC[75]

62. Hang, H.; Steinwart, I. Fast learning from galpha-mixing observations. *J. Multivar. Anal.* **2014**, *127*, 184–199.

63. Fukumizu, K. Kernel Methods for Dependence and Causality. Machine Learning Summer School, Tübingen, Germany, 20–31 August 2007.

64. Shawe-Taylor, J.; Cristianini, N. *Kernel Methods for Pattern Analysis*; Cambridge University Press: New York, NY, USA, 2004.

65. Hamilton, J.D. *Time Series Analysis*; Princeton University Press: Princeton, NJ, USA, 1994.

66. Hatemi-J, A. Asymmetric causality tests with an application. *Empir. Econ.* **2012**, *43*, 447–456.

67. 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: New York, NY, USA, 2003.

68. Pozzi, F.; di Matteo, T.; Aste, T. Spread of risk across financial markets: Better to invest in the peripheries. *Sci. Rep.* **2013**, *3*, 1665.

69. Steinwart, I.; Christmann, A. *Support Vector Machines*; Springer: New York, NY, USA, 2008.

70. A metric space is complete if every Cauchy sequence converges in this space.

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

72. If $\mathcal{F}$ is a normed space, the space of all continuous linear functionals $T : \mathcal{F} \to \mathbb{R}$ is called a topological dual space of $\mathcal{F}$ and denoted as $\mathcal{F}'$.

73. We will require that space F must be separable (to have a complete orthonormal system), but in practice, we will use $\mathbb{R}^n$ and, therefore, that would not be an issue.

74. Pan, A.; Sun, Q.; Bernstein, A.M.; Schulze, M.B.; Manson, J.E.; Stampfer, M.J.; Willett, W.C.; Hu, F.B. Red meat consumption and mortality: Results from 2 prospective cohort studies. *Arch. Internal Med.* **2012**, *172*, 555–563.

75. Chwialkowski, K.; Gretton, A. A Kernel Independence Test for Random Processes. **2014**, *arXiv:1402.4501 [stat]* .