

# Robust Non-linear Wiener-Granger Causality For Large High-dimensional Data

Mehrdad Jafari-Mamaghani<sup>1,2</sup>

April 2014

1. Division of Mathematical Statistics, Department of Mathematics, Stockholm University, Stockholm, Sweden
2. Center for Biosciences, Department of Biosciences and Nutrition, Karolinska Institutet, Huddinge, Sweden

## Abstract

Wiener-Granger causality is a widely used framework of causal analysis for temporally resolved events. We introduce a new measure of Wiener-Granger causality based on kernelization of partial canonical correlation analysis with specific advantages in the context of large high-dimensional data. The introduced measure is able to detect non-linear and non-monotonous signals, is designed to be immune to noise, and offers tunability in terms of computational complexity in its estimations. Furthermore, we show that, under specified conditions, the introduced measure can be regarded as an estimate of conditional mutual information (transfer entropy). The functionality of this measure is assessed using comparative simulations where it outperforms other existing methods. The paper is concluded with an application to climatological data.

KEY WORDS: Wiener-Granger causality, Kernel canonical correlation analysis, High-dimensional learning

## 1 Introduction

Big data is a frequent topic in contemporary discourse on applications of statistics [1]; a topic that is likely to sustain its weight with diminishing costs of data storage and increasing accessibility to digital data. Here, a challenge, or an opportunity, is to move beyond identifying correlations to discovering causations using probability theory and statistics [2]. In statistics, frameworks of causal analysis include the Neyman-Rubin models of causal inference, Bayesian networks, and models of Wiener-Granger causality. The first class of models are designed for experimental static data and thus, inappropriate for large observational data.

Bayesian networks are based on a conceptualization of causal learning on static data using probabilistic graphical models with algorithms operating on estimated (or a priori

known) probability densities [3]. Bayesian networks can be extended to time-resolved domains using *dynamic* Bayesian networks (DBN), a generalization of hidden Markov models [4]. However, prior knowledge of probability densities is an uncommon occurrence in observational data and density estimation is a non-trivial task in high-dimensional settings.

Wiener-Granger causality operates on stochastic processes and measures causality by using the relative reduction in prediction error. As Wiener-Granger causality has been proven to outperform DBN in large data sets [5], and as it additionally offers frequency domain decomposition [5, 6], we argue that Wiener-Granger causality is the most suitable choice for modeling causality in the context of (temporally resolved) large high-dimensional data.

Wiener-Granger causality (WGC), first conceived by Wiener [7], coined and parametrized by Granger [8], and developed further by Sims [9] and Geweke [10], has led to a rich repertoire of analytical methods, conceptual spin-offs, and applications in social and natural sciences. Dominant fields of applications of WGC include econometrics [11, 12, 13], neurophysiology [14, 15, 16], climatology [17, 18, 19], and most recently in cell biology [20]. However, it should be noted that although WGC is not synonymous with causality in a philosophical sense, it offers a solid framework to quantitatively measure and verify a specific type of causality.

WGC conceptualizes the event where the cause temporally precedes the effect, and where the embedded information in the cause about the effect is unique [8]. Employing the nomenclature of probability theory and omitting vector notations without any loss of generality, under the null hypothesis, given  $k$  lags and the temporally resolved random continuous vectors  $A, B$  and the set of all other random vectors  $C$  in any arbitrary system (or probability space  $\Omega$ ),  $\{B\}$  does not Wiener-Granger cause  $\{A\}$ , if:

$$H_0 : A_t \perp\!\!\!\perp \{B_{t-1}, \dots, B_{t-k}\} | \{A_{t-1}, \dots, A_{t-k}, C_{t-1}, \dots, C_{t-k}\}, \quad (1)$$

where  $\perp\!\!\!\perp$  denotes probabilistic independence and where we have assumed stochastic processes that are time-invariant/stationary. For the sake of simplicity, we implement the following substitutions in the remainder of this study:  $X = A_t$ ,  $Y = \{B\}_{t-1}^{t-k}$ , and  $Z = \{A, C\}_{t-1}^{t-k}$ . Accordingly, (1) can be reformulated as:

$$H_0 : X \perp\!\!\!\perp Y | Z, \quad (2)$$

under which [21]:

$$f(X|Z) = f(X|YZ). \quad (3)$$

Under the alternate hypothesis  $H_1 : X \not\perp\!\!\!\perp Y | Z$ , where  $f(X|Z) \neq f(X|YZ)$  we say that  $Y$  (Wiener-Granger) causes  $X$  given  $Z$ .

Moving beyond the preliminaries, the next task is to test the hypothesis in (2). The method proposed by Granger to test this hypothesis is based on linear models and a corresponding variance-based test statistic [8]. More specifically, using linear regression to model the probability density functions in (3), the hypothesis in (2) can be tested using the F-distributed Granger-Sargent test based on restricted and unrestricted residual sums of squares. Succeeding models have extended this concept to the multivariate setting using the *generalized* variance [6, 22] and the *total* variance [23]. Other linear models have

utilized feature selection techniques such as the Lasso in high-dimensional applications of WGC [24, 25]. Naturally, the major drawback of linear models is their insensitivity to non-linear relationships, a common occurrence in empirical data. This shortcoming is often circumvented by the usage of non-linear (but parametric) models of WGC. To name a few, non-linear Fourier and wavelet transformations [26, 27], radial basis functions [28, 29], copula density estimations [30], and additive models [31], have all been employed in non-linear analysis of WGC. Although the choice of model can be a non-trivial issue, the primary disadvantage of non-linear models lies in the risk of overfitting, which usually leads to retrieval of spurious causations.

Another class of models used in analysis of WGC are based on the information theoretical approach and hence non-parametric. Where parametric models are based on the covariance, non-parametric models employ the Shannon entropy [32] to quantify the deviation between the probability density functions in (3). The most well-known measure in this context is the *transfer entropy* [33], equivalent to the conditional mutual information [34, 35, 36, 37]. Interestingly, there is an equivalence between transfer entropy and the linear generalized variance test of WGC for Gaussian variables [38, 39]. Although non-parametric models offer an attractive framework for analysis of WGC, it has been observed that estimates of information theoretical measures suffer from insensitivity to non-linear relationships [40]. This issue is potentially exacerbated in the context of high-dimensional data where estimation of multivariate probability densities, needed to estimate information theoretical measures, is a non-trivial task and requires a fair amount of supervision and tuning.

Canonical correlation analysis (CCA) [41], another statistical tool employed in analysis of WGC, can be seen as an extension of multiple regression to multivariate regression, i.e. reduced rank regression. In fact the solutions to reduced rank regression can be estimated via CCA [42, 43]. Moreover, it has been shown that the solutions from CCA are equivalent to those of orthonormalized partial least squares [44]. Given two multivariate sets of data, CCA finds linear projections of each set such that the between-projection correlations are maximized. Similar to principal component analysis, the projections are ordered based on the eigenvalues of covariance matrices; a property that is of particular interest in the presence of elevated levels of noise [45]. In the context of WGC, CCA has been employed in order to overcome singularities resulting from high-dimensionality of data [46], and partial CCA (PCCA) has been employed in neural connectivity analysis [47], and studies of optical flows in videos [48].

In this paper, we will introduce a new measure of Wiener-Granger causality based on kernel PCCA. More specifically, we will derive a measure of WGC based on kernel partial canonical correlation analysis (KPCCA) in the reproducing kernel Hilbert spaces (RKHS). The issues of overfitting and computational complexity are tackled by penalization and parsimonious matrix decomposition, respectively. In addition, we will show that under certain conditions, the measure based on KPCCA can be regarded as an estimate of transfer entropy (conditional mutual information), a particularly appealing property since estimation of transfer entropy in high-dimensional settings is a non-trivial and expensive task [49]. We will also discuss the similarities and differences between the measure in this paper and closely related measures in the context of WGC.

In the following, Section 2 introduces the methods, Section 3 presents data simulations, climatological data, and corresponding results. Lastly, Section 4 closes the paper

with the concluding remarks.

## 2 Methods

In this section we will build up the framework needed to introduce KPCCA. We will adopt the same variables as those used in the Introduction and follow the hypothesis in (2) unless otherwise stated. Furthermore, without any loss of generality, we assume that all multivariate random vectors in the remainder of this paper are standard Gaussians with  $N$  observations, mean zero and standard deviation 1, and denote the dimension of each random vector by  $d_X$ ,  $d_Y$  and so on.

### 2.1 Canonical correlations

#### 2.1.1 Correlation analysis

The product-moment correlation coefficient between any two random variables  $A$  and  $B$  is defined by:

$$\rho_{A,B} = \frac{\Sigma_{AB}}{\sigma_A \sigma_B} \quad (4)$$

where  $\Sigma_{AB}$  represents the covariance of  $A$  and  $B$ , and  $\sigma_A$  the standard deviation of  $A$ . Following the earlier specification of centered random variables,  $\rho_{A,B}$  can be reformulated as:

$$\rho_{A,B} = \frac{A'B}{\sqrt{A'A}\sqrt{B'B}} \quad (5)$$

$$= \frac{\langle A, B \rangle}{\langle A, A \rangle^{1/2} \langle B, B \rangle^{1/2}} \quad (6)$$

where  $A'$  denotes the transpose of  $A$  in (5), and  $\langle A, A \rangle$  is called the *inner product* of  $A$  in (6). The non-parametric variety of  $\rho$  is the Spearman's rank correlation coefficient  $\rho^{(r)}$  and is simply based on the ranks of  $A$  and  $B$ . Denoting the ranks of  $A$  and  $B$  as  $A^{(r)}$  and  $B^{(r)}$ , respectively, the value of  $\rho^{(r)}$  is determined by plugging in  $A^{(r)}$  and  $B^{(r)}$  in (4).

#### 2.1.2 Canonical correlation analysis

Formulated in 1936 by Hotelling [41], canonical correlation analysis (CCA) extends the concept in the previous section to two (or more) *sets* of variables. Re-employing the random vectors  $X$  and  $Y$ , CCA is designed to find linear projections of  $X$  and  $Y$  subject to the maximization of correlations between these projections. The linear projections yielded by CCA are closely linked to the linear projections in principal component analysis (PCA), and the condition of between-set correlation maximization can be regarded as a generalization of the solution in partial least squares (PLS) [50].

Let  $\alpha$  and  $\beta$  denote linear projections of the multivariate vectors  $X$  and  $Y$ , respectively. The *canonical correlations* of  $X$  and  $Y$  are defined by:

$$\begin{aligned}\rho_{X,Y}^{(\text{CCA})} &= \arg \max_{\alpha, \beta} \rho_{X\alpha, Y\beta} \\ &= \arg \max_{\alpha, \beta} \frac{\langle X\alpha, Y\beta \rangle}{\langle X\alpha, X\alpha \rangle^{1/2} \langle Y\beta, Y\beta \rangle^{1/2}}.\end{aligned}\quad (7)$$

Taking the derivatives of (7) with respect to  $\alpha$  and  $\beta$  yields:

$$\begin{aligned}\frac{\partial}{\partial \alpha} \rho_{X,Y}^{(\text{CCA})} &= \langle X, Y \rangle \beta - \frac{\langle X\alpha, Y\beta \rangle}{\langle X\alpha, X\alpha \rangle} \langle X, X \rangle \alpha \\ \frac{\partial}{\partial \beta} \rho_{X,Y}^{(\text{CCA})} &= \langle Y, X \rangle \alpha - \frac{\langle X\alpha, Y\beta \rangle}{\langle Y\beta, Y\beta \rangle} \langle Y, Y \rangle \beta.\end{aligned}$$

Thus, by setting the derivatives to zero and letting  $\langle X\alpha, X\alpha \rangle = \langle Y\beta, Y\beta \rangle = 1$ , we arrive at the following generalized eigenvalue problem, using which, one can estimate  $\alpha$  and  $\beta$ :

$$\begin{pmatrix} 0 & \langle X, Y \rangle \\ \langle Y, X \rangle & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \rho_{X,Y}^{(\text{CCA})} \begin{pmatrix} \langle X, X \rangle & 0 \\ 0 & \langle Y, Y \rangle \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}.\quad (8)$$

The eigenvalues of the system in (8) lead to  $d = \min(d_X, d_Y)$  canonical correlations. The solution above can be extended to more than two sets of variables as demonstrated in [51, 52].

### 2.1.3 Partial canonical correlation analysis

The first reformulation of CCA to include conditionality was by Rao in 1969 where the canonical correlations of  $X$  and  $Y$  conditioned on a third vector  $Z$  are retrieved by removing the influence of  $Z$  using regression techniques [53]. Here we pursue a similar concept to introduce partial canonical correlation analysis (PCCA) based on *partial* covariances.

Let  $\rho_{X,Y|Z}^{(\text{PCCA})}$  denote the partial canonical correlations of  $X$  and  $Y$  given the information in  $Z$ :

$$\rho_{X,Y|Z}^{(\text{PCCA})} = \arg \max_{\alpha, \beta} \frac{\langle (X|Z)\alpha, (Y|Z)\beta \rangle}{\langle (X|Z)\alpha, (X|Z)\alpha \rangle^{1/2} \langle (Y|Z)\beta, (Y|Z)\beta \rangle^{1/2}}$$

where  $X|Z$  denotes the random vector  $X$  with discarded influence of  $Z$ . Using a scheme similar to the previous section, the partial canonical correlations are determined by the following generalized eigenvalue problem:

$$\begin{aligned}& \begin{pmatrix} 0 & \langle X, Y \rangle - \langle X, Z \rangle \langle Z, Z \rangle^{-1} \langle Z, Y \rangle \\ \langle Y, X \rangle - \langle Y, Z \rangle \langle Z, Z \rangle^{-1} \langle Z, X \rangle & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \\ &= \rho_{X,Y|Z}^{(\text{PCCA})} \begin{pmatrix} \langle X, X \rangle - \langle X, Z \rangle \langle Z, Z \rangle^{-1} \langle Z, X \rangle & 0 \\ 0 & \langle Y, Y \rangle - \langle Y, Z \rangle \langle Z, Z \rangle^{-1} \langle Z, Y \rangle \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}\end{aligned}\quad (9)$$

where the non-zero elements in the quadratic matrices are partial covariances.

### 2.1.4 Kernel partial canonical correlation analysis

Kernel methods have had a substantial impact in statistical learning techniques in the past decade [4, 43, 54]. Kernel methods accommodate the analysis of non-linear phenomena using linear techniques by mapping data from an input space into a higher-dimensional feature space  $\mathcal{F}$ :

$$\Xi : X_1, \dots, X_{d_X} \mapsto \Xi(X_1), \dots, \Xi(X_{d_X}).$$

In the context of large data sets, *Mercer kernels* are of particular utility. Let  $X$  be defined on the space  $\mathcal{X}$ , and let  $\kappa(X_i, X_j)$  be a function from  $\mathcal{X} \times \mathcal{X}$  to  $\mathbb{R}$ . If the matrix  $K$ , defined element-wise as  $K_{ij} = \kappa(X_i, X_j)$ , is a positive semidefinite matrix, the function  $\kappa(X_i, X_j)$  is a *Mercer kernel*, and the matrix  $K$  a *Gram matrix*. Under these conditions there is a space  $\mathcal{F}$  and a map  $\Xi$  such that  $\kappa(X_i, X_j)$  is the inner product in  $\mathcal{F}$  between the images  $\Xi(X_i)$  and  $\Xi(X_j)$ :

$$\langle \Xi(X), \Xi(X) \rangle = \kappa(X, X) = K_{X,X}.$$

A continuous kernel qualifying as a Mercer kernel is the Gaussian kernel:  $\Phi(X) = \exp(-X^2/2)/\sqrt{2\pi}$ , which in addition associates  $\kappa(\cdot, \cdot)$  with a *reproducing kernel Hilbert space* (RKHS) [54]. The appealing aspect of the theory above lies within the properties of the Gram matrix  $K$ : under the Mercer conditions,  $K$  can conveniently be factorized to lower-dimensional matrices to counterbalance computational complexity (see below).

The usage of kernel methods in analysis of WGC is not an uncommon practice. A potent approach is presented in [55] where using statistical tests, the Gram matrices used in non-linear analysis of WGC are pruned to circumvent overfitting. More closely related to the topic of this paper, a kernelization of CCA in analysis of WGC is presented in [56]. However, the presented framework here is not immune to the likely issues of overfitting and overcomplexity (cf. the sections below).

Assuming centered Gram matrices in the remainder, the kernel partial canonical correlations of  $X$  and  $Y$  given the information in  $Z$  are defined by:

$$\rho_{X,Y|Z}^{(\text{KPCCA})} = \arg \max_{\gamma, \delta} \frac{\langle \Phi(X|Z)\gamma, \Phi(Y|Z)\delta \rangle}{\langle \Phi(X|Z)\gamma, \Phi(X|Z)\gamma \rangle^{1/2} \langle \Phi(Y|Z)\delta, \Phi(Y|Z)\delta \rangle^{1/2}} \quad (10)$$

where  $\Phi(X|Z)$  denotes the image of the random vector  $X$  in the feature space  $\mathcal{F}$  with removed influence of  $Z$ .

The kernelization of PCCA as conducted above can lead to two issues in real-world applications. Firstly, the naive kernelization in (10) can yield biased canonical correlations. That is, due to the potential risk of overfitting, the kernel-defined spaces spanned by  $\langle \Phi(X|Z), \Phi(X|Z) \rangle$  and  $\langle \Phi(Y|Z), \Phi(Y|Z) \rangle$  are likely to be oriented in identical directions and result in perfect canonical correlations [52]. Secondly, the parametrization in (10) requires the evaluation of full-rank Gram matrices, which is likely to lead to a prohibitive computational expenditure for large data samples.

In the following, the two issues of overfitting and computational efficiency are addressed using regularization and incomplete Cholesky decomposition, respectively.

### 2.1.5 Regularization

The issue of overfitting, leading to falsely high correlations, can be circumvented by regularization. We use the partial least squares norm penalization with regularization parameter  $\zeta$  to reformulate  $\rho_{X,Y|Z}^{(\text{KPCCA})}$ :

$$\rho_{X,Y|Z}^{(\text{KPCCA})} = \arg \max_{\gamma, \delta} \frac{\langle \Phi(X|Z)\gamma, \Phi(Y|Z)\delta \rangle}{\Psi(\Phi(X|Z)\gamma, \zeta)^{1/2} \Psi(\Phi(Y|Z)\delta, \zeta)^{1/2}} \quad (11)$$

where

$$\Psi(\Phi(X|Z)\gamma, \zeta) = \langle \Phi(X|Z)\gamma, \Phi(X|Z)\gamma \rangle + \zeta \|\langle \gamma, \Phi(X|Z)\gamma \rangle\|^2, \quad (12)$$

$$\Psi(\Phi(Y|Z)\delta, \zeta) = \langle \Phi(Y|Z)\delta, \Phi(Y|Z)\delta \rangle + \zeta \|\langle \delta, \Phi(Y|Z)\delta \rangle\|^2. \quad (13)$$

By expanding the terms in (12) and (13) up to the second order in  $\zeta$  and deriving the derivatives of (11), we arrive at the following generalized eigenvalue problem:

$$\begin{pmatrix} 0 & K_{X,Y|Z} \\ K_{Y,X|Z} & 0 \end{pmatrix} \begin{pmatrix} \gamma \\ \delta \end{pmatrix} = \rho_{X,Y|Z}^{(\text{KPCCA})} \begin{pmatrix} K_{X,X|Z} + \zeta I & 0 \\ 0 & K_{Y,Y|Z} + \zeta I \end{pmatrix} \begin{pmatrix} \gamma \\ \delta \end{pmatrix} \quad (14)$$

where  $I$  is the identity matrix and

$$K_{X,Y|Z} = \langle \Phi(X), \Phi(Y) \rangle - \langle \Phi(X), \Phi(Z) \rangle \langle \Phi(Z), \Phi(Z) \rangle^{-1} \langle \Phi(Z), \Phi(Y) \rangle.$$

The solution obtained by solving (14) is similar to the solution yielded by the *Canonical Ridge* [48, 57].

### 2.1.6 Incomplete Cholesky decomposition

Given a positive semidefinite  $d \times d$  matrix  $K$ , Cholesky decomposition finds a factorization of  $K$  such that  $K = GG'$ , where  $G$  is a lower triangular  $d \times d$  matrix, and  $G'$  denotes the transpose of  $G$ . Our aim is to approximate  $G$  using a low-rank  $d \times c$  matrix  $\tilde{G}$  where  $c \ll d$ , and the difference  $K - \tilde{G}\tilde{G}'$  has a norm less than a given threshold  $\tau$ . This aim is achieved by incomplete Cholesky decomposition [58]. Whereas complete Cholesky decomposition employs all pivots in factorization, incomplete Cholesky decomposition skips those pivots that are below a given level. In fact, incomplete Cholesky decomposition is the dual implementation of partial Gram-Schmidt orthogonalization [52]. The complete algorithm for incomplete Cholesky decomposition can be found in [52].

## 2.2 Measures of causality

Using the same random vectors as in (2), and the definition of partial canonical correlations above, we define the measure *Canonical Wiener-Granger Causality* (CC) as:

$$\mathcal{C}_{Y \rightarrow X|Z} := -\frac{1}{2} \ln \prod_{i=1}^d \left[ 1 - \left( \rho_{X,Y|Z}^{(\text{PCCA})}(i) \right)^2 \right] \quad (15)$$

where  $\rho_{X,Y|Z}^{(\text{PCCA})}(i)$  denotes the  $i$ th partial canonical correlation, and  $d = \min(d_X, d_Y)$ . As it turns out, when  $d = d_X = d_Y$  for Gaussian variables, the measure in (15) is equivalent to the transfer entropy of  $X$  and  $Y$  given  $Z$ :

$$\mathcal{C}_{Y \rightarrow X|Z} = \mathcal{T}_{Y \rightarrow X|Z}. \quad (16)$$

A proof of (16) is given in A.1. As a consequence, under the specified conditions, the maximum likelihood estimator of (15) follows a  $\chi^2$  distribution under the null hypothesis (asymptotically for large samples). In any other case, tests of significance can be carried out using permutation resampling.

Similarly, we define the measure *Kernel Canonical Wiener-Granger Causality* (KCC) as:

$$\mathcal{C}_{Y \rightarrow X|Z}^{\mathcal{F}} := -\frac{1}{2} \ln \prod_{i=1}^{d(\mathcal{F})} \left[ 1 - \left( \rho_{X,Y|Z}^{(\text{KPCCA})}(i) \right)^2 \right]. \quad (17)$$

Accordingly, given Gaussian distributed images of input data in the higher-dimensional feature space  $\mathcal{F}$ , where  $d(\mathcal{F}) = d(\mathcal{F}_X) = d(\mathcal{F}_Y)$ , we can show that  $\mathcal{C}_{Y \rightarrow X|Z}^{\mathcal{F}} = \mathcal{T}_{Y \rightarrow X|Z}^{\mathcal{F}}$ , where the term on the right hand side denotes the transfer entropy of  $\Phi(Y)$  to  $\Phi(X)$  given  $\Phi(Z)$  in the feature space  $\mathcal{F}$ . The proof follows by analogy from A.1.

Although the equality in (16) provides a useful estimate of transfer entropy using PCCA, the assumption of Gaussianity for observed processes is not always met in real-world applications. However, using the Gaussian kernel  $\Phi(X) = \exp(-X^2/2)/\sqrt{2\pi}$ , we argue that:

$$\mathcal{C}_{Y \rightarrow X|Z}^{\mathcal{F}} = \mathcal{T}_{Y \rightarrow X|Z}. \quad (18)$$

That is, the KCC based on images of data in the feature space  $\mathcal{F}$  derived via Gaussian kernels, coincides with the transfer entropy of data in the input space. This result follows from the relation between kernel generalized variance and mutual information. Further remarks underlying the equality in (18) are given in A.2.

It follows from (18) that:

$$\mathcal{T}_{Y \rightarrow X|Z}^{\mathcal{F}} = \mathcal{T}_{Y \rightarrow X|Z}, \quad (19)$$

leading to the conclusion that the transfer of information from  $Y$  to  $X$  given  $Z$  is identical regardless of its evaluation in the input or the feature space. The intuition here is underpinned by the fact that kernelization of data only serves to make information more accessible rather than altering its content.

The utilities of KCC as a measure of causality can be grouped into: *i*) its innate properties, and *ii*) its advantage over estimates of entropy-based measures. Firstly, KCC is preferable to linear measures as it is capable of detecting non-linear causal signals. Furthermore, as we shall see in applications to synthetic data, by using the eigenvalues of Gram matrices, KCC is founded only on the canonical information, disregarding spurious signals. This feature is highly relevant in the context of high-dimensional input spaces where the pervasiveness of noise may lead to difficulties in discovering the underlying relationships. Additionally, by employing incomplete Cholesky decomposition, we are able to control the computational complexity of the main bottleneck in estimations of KCC.



Secondly, KCC provides a superior alternative to estimates of differential/continuous transfer entropy based on kernel density estimation (KDE) [59, 60] as these estimates, in the context of estimating information theoretical measures, suffer from insensitivity to non-linear relationships [40, 36]. Moreover, KDE-based estimates need supervision in terms of bandwidth selection and demand substantial computational resources in high-dimensional settings.

In the following, we will assess the performance of CC and KCC using synthetic and real-world data.

## 3 Results

### 3.1 Synthetic data

To assess the functionality of the proposed measures of causality, CC and KCC, we will use the system in (20) to simulate synthetic multivariate time-series exhibiting non-monotonous causal signals. We define the 20-dimensional multivariate vectors  $X, Y^{(1)}, Y^{(2)}, Y^{(3)}$  and  $Y^{(4)}$  as:

$$\begin{cases} X_t = s, & s \sim \mathcal{N}(0, I) \\ X_{t-i} = s + 0.1 \cdot i \cdot \epsilon, & i = 1, \dots, k, \quad \epsilon \sim \mathcal{N}(0, I) \\ \begin{cases} Y_t^{(1)} = \epsilon & \epsilon \sim \mathcal{N}(0, I) \\ Y_{t-i}^{(1)} = \sin(5X_t) + 0.1 \cdot i \cdot \epsilon, & i = 1, \dots, k \end{cases} \\ \begin{cases} Y_{t-i}^{(2)} = \epsilon, & i = 0, 1 \\ Y_{t-j}^{(2)} = \log[\text{abs}(X_t + 0.25 \cdot (j-1) \cdot \epsilon)], & j = 2, \dots, k \end{cases} \\ \begin{cases} Y_{t-i}^{(3)} = \epsilon, & i = 0, \dots, 2 \\ Y_{t-j}^{(3)} = (-1)^\omega \cdot X_t + 0.2 \cdot (j-2) \cdot \epsilon, & j = 3, \dots, k, \quad \omega \sim \mathcal{U}(1, 2) \end{cases} \\ \begin{cases} Y_{t-i}^{(4)} = \epsilon, & i = 0, \dots, 3 \\ Y_{t-j}^{(4)} = \exp(X_t) + 1.3 \cdot (j-3) \cdot \epsilon, & j = 4, \dots, k \end{cases} \end{cases} \quad (20)$$

where  $\mathcal{N}(0, I)$  denotes a multivariate standard Gaussian, and  $\mathcal{U}(a, b)$  is a discrete uniform distribution. Density plots in Figure 1 provide a display of the non-linear relationships in (20). In our simulations we have used  $k = 4$  lags in 10,000 realizations of the system above. The results, as a comparison between CC, KCC, GenVar (the generalized variance test, see (24)), and transfer entropy/conditional mutual information as estimated in [36], using 1000 permutation resamplings, are presented in Figure 2. In applications of KCC to the system above, the width of the Gaussian kernel has been set to  $\sigma = 1$  after standardization of data, the penalization parameter  $\zeta = 10^{-7}$ , and the incomplete decomposition of kernel matrices uses the threshold  $\tau = 10^{-6}$ . In Figure 2, the presence of an arrow from, e.g.,  $Y^{(1)}$  to  $X$  at lag  $k = 2$ , denotes the rejection of the hypothesis:

$$H_0 : X_t \perp\!\!\!\perp Y_{t-1}^{(1)}, Y_{t-2}^{(1)} \mid X_{t-1}, X_{t-2}, Y_{t-1}^{(2)}, Y_{t-2}^{(2)}, Y_{t-1}^{(3)}, Y_{t-2}^{(3)}, Y_{t-1}^{(4)}, Y_{t-2}^{(4)}.$$

From the networks in Figure 2, it is apparent that GenVar and CC fail to discover the non-linear and non-monotonous relationships. However, generalized variance and CC do

capture the exponential causal signal from  $Y^{(4)}$  to  $X$  due to its monotonicity. Transfer entropy succeeds only in detecting the causal signals from  $Y^{(1)}$  and  $Y^{(2)}$  to  $X$  for up to  $k = 3$  lags and thus, fails to find the remaining signals due to increased dimensionality and levels of noise. In contrast, KCC successfully discovers all the planted relationships absent of any spurious causations.

Moreover, to extend the scope of analysis based on the performance of KCC, 100 realizations of the system in (20) were performed to outline the robustness of KCC in terms of rate of discovery of causal links, and coefficients of variation of KCC scores (see Figure 3). The results present unanimously low false discovery rates ( $< 3\%$ ), and uniformly consistent discovery of the planted signals.

### 3.2 Climatological data

To test the performance of KCC on real-world large high-dimensional data, an analysis of WGC is performed on gridded climatological records. It should be added that the aim here is not to draw any exclusive conclusions on climatological mechanisms, but to test the utility of KCC on types of large datasets that occur with increasing frequency.

The data contains information on temperature and precipitation from December 1765 to November 2000, recorded in the North Atlantic/European region ( $80\text{-}30^\circ\text{N}$  and  $50^\circ\text{W}\text{-}40^\circ\text{E}$ ) [61]; see Figure 4. As climatological data reconstructed from proxies often suffers from relatively higher levels of noise than instrumental data [62], the analysis here is limited to the period 1900-2000 based on actual records. Furthermore, the 20th century is divided into the two periods 1900-1960 and 1961-2000 due to evidence of increased global warming effects in the latter period [61]. Lastly, to ensure stationarity, i.e. avoid seasonal variations, we confine the analysis to a specific month.

In the applied analysis herein, we investigate whether the variation in temperature and precipitation in the North Atlantic regions of Europe (designation adapted from the European Environmental Agency, the black-crossed regions in Figure 4), is causal of temperature in other regions of Europe in the month April:

$$H_0 : T_{April}^{(x)} \perp\!\!\!\perp T_{March}^{(NA)}, P_{March}^{(NA)} \mid T_{March}^{(x)}, P_{March}^{(x)}. \quad (21)$$

In the hypothesis above  $T$  denotes temperature,  $P$  denotes precipitation, and the superscript  $x$  denotes any region outside of the North Atlantic (NA) region (here, we regard every pixel in the gridded records outside of the NA region as an independent region). The results, based on tests of the hypothesis in (21), for the two periods 1900-1960 and 1961-2000 are displayed in Figure 4. The application of KCC here uses the same parameters as in the previous section after data standardization.

It is seen that the variation of temperature and precipitation in the North Atlantic regions of Europe in March has a stronger causal effect on the April temperatures in the continental regions in the latter parts of the 20th century. In other words, in the period 1960-2000, temperature and precipitation in the North Atlantic regions in March contain unique information about the temperatures of continental Europe in April; a phenomenon that does not seem to be present in the earlier parts of the 20th century.

## 4 Discussion

Wiener-Granger causality is becoming a routine practice in causal analysis of temporally resolved observational data. In the context of large high-dimensional data, the issues of non-linearity, noise pervasiveness, and computational complexity present a series of challenges to classical frameworks of analysis of Wiener-Granger causality. The aim of this paper has been to present an alternative approach to measure Wiener-Granger causality, which circumvents all of the three issues above.

The proposed measure of Wiener-Granger causality (KCC) in this paper based on kernel partial canonical correlations and parsimonious matrix factorization offers the following advantages: *i)* as demonstrated by the simulations, KCC is sensitive to non-linear and non-monotonous signals absent of any display of spurious causalities, *ii)* as seen in the comparison between transfer entropy and KCC, by using the canonical information, KCC is more immune to noise and high-dimensionality, *iii)* when measured using Mercer kernels, KCC offers appealing tunability in terms of computational complexity of its main estimation bottleneck and is capable of operating considerably faster than similar kernel methods, and *iv)* most importantly, when measured using Gaussian kernels, KCC can be regarded as an estimate of transfer entropy/conditional mutual information.

All of these properties offer clear advantages over currently existing measures in non-linear/non-parametric analysis of Wiener-Granger causality. Given this background, we believe that KCC has the potential to become a standard measure as a result of its superior performance, computational feasibility, and straightforward implementation. Future work on this topic includes closer investigation of KCC and its relation to other existing measures and more comprehensive applications to real-world data.

## Acknowledgements

The author wishes to thank John Hertz, Joanna Tyrcha and John G. Lock for insightful comments. The author has been supported by the Swedish Research Council grant # 340-2012-6011.

## A Appendix

### A.1 Equivalence between CC and transfer entropy

Assume that  $d = d_X = d_Y$  for normalized random vectors  $X$  and  $Y$ . Consider the reformulation of the generalized eigenvalue problem in (9):

$$\begin{pmatrix} \Sigma_{X,X|Z} & \Sigma_{X,Y|Z} \\ \Sigma_{Y,X|Z} & \Sigma_{Y,Y|Z} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = (1 + \rho_{X,Y|Z}^{(\text{PCCA})}) \begin{pmatrix} \Sigma_{X,X|Z} & 0 \\ 0 & \Sigma_{Y,Y|Z} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \quad (22)$$

with eigenvalues  $\{1 + \rho_{X,Y|Z}^{(\text{PCCA})}(1), 1 - \rho_{X,Y|Z}^{(\text{PCCA})}(1), \dots, 1 + \rho_{X,Y|Z}^{(\text{PCCA})}(d), 1 - \rho_{X,Y|Z}^{(\text{PCCA})}(d), 1, \dots, 1\}$ , and where:

$$\Sigma_{X,Y|Z} = \langle X, Y \rangle - \langle X, Z \rangle \langle Z, Z \rangle^{-1} \langle Z, Y \rangle$$

For an invertible matrix  $D$ , the eigenvalues of the generalized eigenvalue problems  $Cx = (1 + \rho)Dx$  and  $D^{-1}Cx = (1 + \rho)x$  are identical. As a result,  $\mathcal{C}_{Y \rightarrow X|Z}$  can be expressed as:

$$\begin{aligned}
\mathcal{C}_{Y \rightarrow X|Z} &= -\frac{1}{2} \ln \prod_{i=1}^d \left[ 1 - \rho_{X,Y|Z}^{(\text{PCCA})}(i)^2 \right] \\
&= -\frac{1}{2} \ln \prod_{i=1}^d \left[ 1 - \rho_{X,Y|Z}^{(\text{PCCA})}(i) \right] \left[ 1 + \rho_{X,Y|Z}^{(\text{PCCA})}(i) \right] \\
&= -\frac{1}{2} \ln \left| \begin{pmatrix} \Sigma_{X,X|Z} & 0 \\ 0 & \Sigma_{Y,Y|Z} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{X,X|Z} & \Sigma_{X,Y|Z} \\ \Sigma_{Y,X|Z} & \Sigma_{Y,Y|Z} \end{pmatrix} \right| \\
&= -\frac{1}{2} \ln \left( \left| \begin{pmatrix} \Sigma_{X,X|Z} & 0 \\ 0 & \Sigma_{Y,Y|Z} \end{pmatrix} \right|^{-1} \left| \begin{pmatrix} \Sigma_{X,X|Z} & \Sigma_{X,Y|Z} \\ \Sigma_{Y,X|Z} & \Sigma_{Y,Y|Z} \end{pmatrix} \right| \right) \\
&= \frac{1}{2} \ln |\Sigma_{X,X|Z}| + \frac{1}{2} \ln |\Sigma_{Y,Y|Z}| - \frac{1}{2} \ln \left| \begin{pmatrix} \Sigma_{X,X|Z} & \Sigma_{X,Y|Z} \\ \Sigma_{Y,X|Z} & \Sigma_{Y,Y|Z} \end{pmatrix} \right| \\
&= H(X|Z) + H(Y|Z) - H(X, Y|Z) \\
&= I(X, Y|Z) = \mathcal{T}_{Y \rightarrow X|Z}.
\end{aligned} \tag{23}$$

Indeed, the transfer entropy  $\mathcal{T}_{Y \rightarrow X|Z}$  is identical to conditional mutual information and the Jensen-Shannon divergence [35, 36]. Due to [38], and under the specified conditions, it also follows that:

$$\begin{aligned}
\mathcal{C}_{Y \rightarrow X|Z} &= \frac{1}{2} \mathcal{G}_{Y \rightarrow X|Z} \\
&= \frac{1}{2} \ln \left( \frac{|\Sigma_{X|Z}|}{|\Sigma_{X|YZ}|} \right)
\end{aligned} \tag{24}$$

where the quantity on the right hand side, the generalized variance test, corresponds to the test statistic proposed by Geweke in [6]. When  $d_X \neq d_Y$  the following relationship is obtained [50]:

$$-\frac{1}{2} \ln \left[ 1 - \max \left( \rho_{X,Y|Z}^{(\text{PCCA})} \right)^2 \right] \leq \mathcal{T}_{Y \rightarrow X|Z} \leq -\frac{\min(d_X, d_Y)}{2} \ln \left[ 1 - \max \left( \rho_{X,Y|Z}^{(\text{PCCA})} \right)^2 \right]$$

where we have only used the maximum canonical correlation.

## A.2 Link between KCC and transfer entropy

The link between KCC and transfer entropy stems from the link between kernel generalized variance and mutual information. Kernel generalized variance approaches a limit as the kernel width approaches zero. In the bivariate case, this limit is equal to the mutual information, up to second order, expanding around independence. A sketch of a proof underlining this property is presented in [50]. This sketch is based on three main components that can easily be extended to establish a link between KCC and transfer entropy. To avoid redundancy, we refer the interested reader to the thorough sketch presented in the mentioned study.

## References

- [1] David Walker and Kaiser Fung. Big data and big business: Should statisticians join in? *Significance*, 10(4):20–25, 2013.
- [2] Kaiser Fung. *Numbersense: How to Use Big Data to Your Advantage*. McGraw Hill Professional, 2013.
- [3] Judea Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, New York, NY, USA, 2000.
- [4] Kevin P Murphy. *Machine learning: a probabilistic perspective*. The MIT Press, 2012.
- [5] Cunlu Zou and Jianfeng Feng. Granger causality vs. dynamic bayesian network inference: a comparative study. *BMC bioinformatics*, 10(1):122, 2009.
- [6] John Geweke. Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association*, 77(378):304, 1982.
- [7] Norbert Wiener. The theory of prediction. *Modern mathematics for engineers*, pages 165–190, 1956.
- [8] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–38, 1969.
- [9] Christopher A Sims. Money, Income, and Causality. *Amer Econ Rev*, 62(4):540–552, 1972.
- [10] John F Geweke. Measures of conditional linear dependence and feedback between time series. *Journal of the American Statistical Association*, 79(388):907–915, 1984.
- [11] Mohammad Reza Lotfalipour, Mohammad Ali Falahi, and Malihe Ashena. Economic growth, co2 emissions, and fossil fuels consumption in iran. *Energy*, 35(12):5115–5120, 2010.
- [12] Douglas M Walker and Peter T Calcagno. Casinos and political corruption in the united states: a granger causality analysis. *Applied Economics*, 45(34):4781–4795, 2013.
- [13] Eric S Lin and Hamid E Ali. Military spending and inequality: panel granger causality test. *Journal of Peace Research*, 46(5):671–685, 2009.
- [14] Klaas E Stephan and Alard Roebroeck. A short history of causal modeling of fMRI data. *NeuroImage*, January 2012.
- [15] J Paul Hamilton, Gang Chen, Moriah E Thomason, Mirra E Schwartz, and Ian H Gotlib. Investigating neural primacy in major depressive disorder: multivariate granger causality analysis of resting-state fmri time-series data. *Molecular psychiatry*, 16(7):763–772, 2010.

- [16] Demian Battaglia, Annette Witt, Fred Wolf, and Theo Geisel. Dynamic effective connectivity of inter-areal brain circuits. *PLoS computational biology*, 8(3):e1002438, 2012.
- [17] Timothy J Mosedale, David B Stephenson, Matthew Collins, and Terence C Mills. Granger causality of coupled climate processes: Ocean feedback on the north atlantic oscillation. *Journal of climate*, 19(7):1182–1194, 2006.
- [18] James B Elsner. Evidence in support of the climate change–atlantic hurricane hypothesis. *Geophysical Research Letters*, 33(16):L16705, 2006.
- [19] Jian Kang and Rolf Larsson. What is the link between temperature and carbon dioxide levels? a granger causality analysis based on ice core data. *Theoretical and Applied Climatology*, pages 1–12, 2013.
- [20] John G Lock, Mehrdad Jafari-Mamaghani, Hamdah Shafqat-Abbasi, Xiaowei Gong, Joanna Tyrcha, and Staffan Strömblad. Plasticity in the macromolecular-scale causal networks of cell migration. *PloS one*, 9(2):e90593, 2014.
- [21] Jean-Pierre Florens and Michel Mouchart. A note on noncausality. *Econometrica: Journal of the Econometric Society*, pages 583–591, 1982.
- [22] Adam B Barrett, Lionel Barnett, and Anil K Seth. Multivariate granger causality and generalized variance. *Physical Review E*, 81(4):041907, 2010.
- [23] Christophe Ladroue, Shuixia Guo, Keith Kendrick, and Jianfeng Feng. Beyond element-wise interactions: identifying complex interactions in biological processes. *PloS one*, 4(9):e6899, 2009.
- [24] Pedro A Valdés-Sosa, Jose M Sánchez-Bornot, Agustín Lage-Castellanos, Mayrim Vega-Hernández, Jorge Bosch-Bayard, Lester Melie-García, and Erick Canales-Rodríguez. Estimating brain functional connectivity with sparse multivariate autoregression. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1457):969–981, 2005.
- [25] Ali Shojaie and George Michailidis. Discovering graphical granger causality using the truncating lasso penalty. *Bioinformatics*, 26(18):i517–i523, 2010.
- [26] Mukeshwar Dhamala, Govindan Rangarajan, and Mingzhou Ding. Estimating granger causality from fourier and wavelet transforms of time series data. *Physical Review Letters*, 100(1):018701, 2008.
- [27] François Benhmad. Modeling nonlinear granger causality between the oil price and us dollar: A wavelet based approach. *Economic Modelling*, 29(4):1505–1514, 2012.
- [28] Nicola Ancona, Daniele Marinazzo, and Sebastiano Stramaglia. Radial basis function approach to nonlinear granger causality of time series. *Phys. Rev. E*, 70:056221, Nov 2004.

- [29] Daniele Marinazzo, Mario Pellicoro, and Sebastiano Stramaglia. Nonlinear parametric model for granger causality of time series. *Phys. Rev. E*, 73:066216, Jun 2006.
- [30] Abderrahim Taamouti, Taoufik Bouezmarni, and Anouar El Gouch. Nonparametric estimation and inference for granger causality measures. 2012.
- [31] Mehmet Balcilar, Zeynel Abidin Ozdemir, and Esin Cakan. On the nonlinear causality between inflation and inflation uncertainty in the g3 countries. *Journal of Applied Economics*, 14(2):269–296, 2011.
- [32] Claude E Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, July 1948.
- [33] Thomas Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85:461–464, Jul 2000.
- [34] Katerina Hlaváčková-Schindler, Milan Paluš, Martin Vejmelka, and Joydeep Bhattacharya. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441(1):1–46, 2007.
- [35] Abd-Krim K Seghouane and Shun-Ichi Amari. Identification of directed influence: granger causality, kullback-leibler divergence, and complexity. *Neural computation*, 24(7):1722–1739, July 2012.
- [36] Mehrdad Jafari-Mamaghani. Non-parametric analysis of granger causality using local measures of divergence. *Applied Mathematical Sciences*, 7(83):4107–4136, 2013.
- [37] Mehrdad Jafari-Mamaghani. Non-parametric wiener-granger causality in partially observed systems. *Norbert Wiener in the 21st Century (21CW), 2014 IEEE Conference*, 10.1109/NORBERT.2014.6893936:1–5, 2014.
- [38] Lionel Barnett, Adam B Barrett, and Anil K Seth. Granger causality and transfer entropy are equivalent for gaussian variables. *Phys. Rev. Lett.*, 103:238701, Dec 2009.
- [39] Mehrdad Jafari-Mamaghani and Joanna Tyrcha. Transfer entropy expressions for a class of non-gaussian distributions. *Entropy*, 16(3):1743–1755, 2014.
- [40] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review. E*, 69(6 Pt 2), June 2004.
- [41] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.
- [42] MK-S Tso. Reduced-rank regression and canonical analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 183–189, 1981.
- [43] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, corrected edition, August 2009.

- [44] Liang Sun, Shuiwang Ji, Shipeng Yu, and Jieping Ye. On the equivalence between canonical correlation analysis and orthonormalized partial least squares. In *IJCAI*, pages 1230–1235, 2009.
- [45] Xiaotong Wen, Govindan Rangarajan, and Mingzhou Ding. Is granger causality a viable technique for analyzing fmri data? *PloS one*, 8(7):e67428, 2013.
- [46] Csilla Horvath, Peter SH Leeftang, and Pieter W Otter. Canonical correlation analysis and wiener-granger causality tests: Useful tools for the specification of var models. *Marketing Letters*, 13(1):53–66, 2002.
- [47] Guo Rong Wu, Fuyong Chen, Dezhi Kang, Xiangyang Zhang, Daniele Marinazzo, and Huaifu Chen. Multiscale causal connectivity analysis by canonical correlation: theory and application to epileptic brain. *Biomedical Engineering, IEEE Transactions on*, 58(11):3088–3096, 2011.
- [48] Yuya Yamashita, Tatsuya Harada, and Yasuo Kuniyoshi. Causal flow. In *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, pages 1–6. IEEE, 2011.
- [49] Anna Zaremba and Tomaso Aste. Measures of causality in complex datasets with application to financial data. *arXiv preprint arXiv:1401.1457*, 2014.
- [50] Francis R Bach and Michael I Jordan. Kernel independent component analysis. *The Journal of Machine Learning Research*, 3:1–48, 2003.
- [51] Jon R Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.
- [52] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.
- [53] B Raja Rao. Partial canonical correlations. *Trabajos de estadística y de investigación operativa*, 20(2):211–219, 1969.
- [54] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels*. The MIT Press, 2002.
- [55] Daniele Marinazzo, Mario Pellicoro, and Sebastiano Stramaglia. Kernel method for nonlinear granger causality. *Physical Review Letters*, 100(14):144103, 2008.
- [56] Guorong Wu, Xujun Duan, Wei Liao, Qing Gao, and Huaifu Chen. Kernel canonical-correlation granger causality for multiple time series. *Physical Review E*, 83(4):041921, 2011.
- [57] Hrishikesh D Vinod. Canonical ridge and econometrics of joint production. *Journal of Econometrics*, 4(2):147–166, 1976.
- [58] Shai Fine and Katya Scheinberg. Efficient svm training using low-rank kernel representations. *The Journal of Machine Learning Research*, 2:243–264, 2002.



- [59] Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.
- [60] Matt P Wand and M Chris Jones. *Kernel smoothing*, volume 60. Crc Press, 1994.
- [61] Carlo Casty, Christoph C Raible, Thomas F Stocker, Heinz Wanner, and Jürg Luterbacher. A european pattern climatology 1766–2000. *Climate Dynamics*, 29(7-8):791–805, 2007.
- [62] Anders Moberg and Gudrun Brattström. Prediction intervals for climate reconstructions with autocorrelated noisean analysis of ordinary least squares and measurement error methods. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 308(3):313–329, 2011.

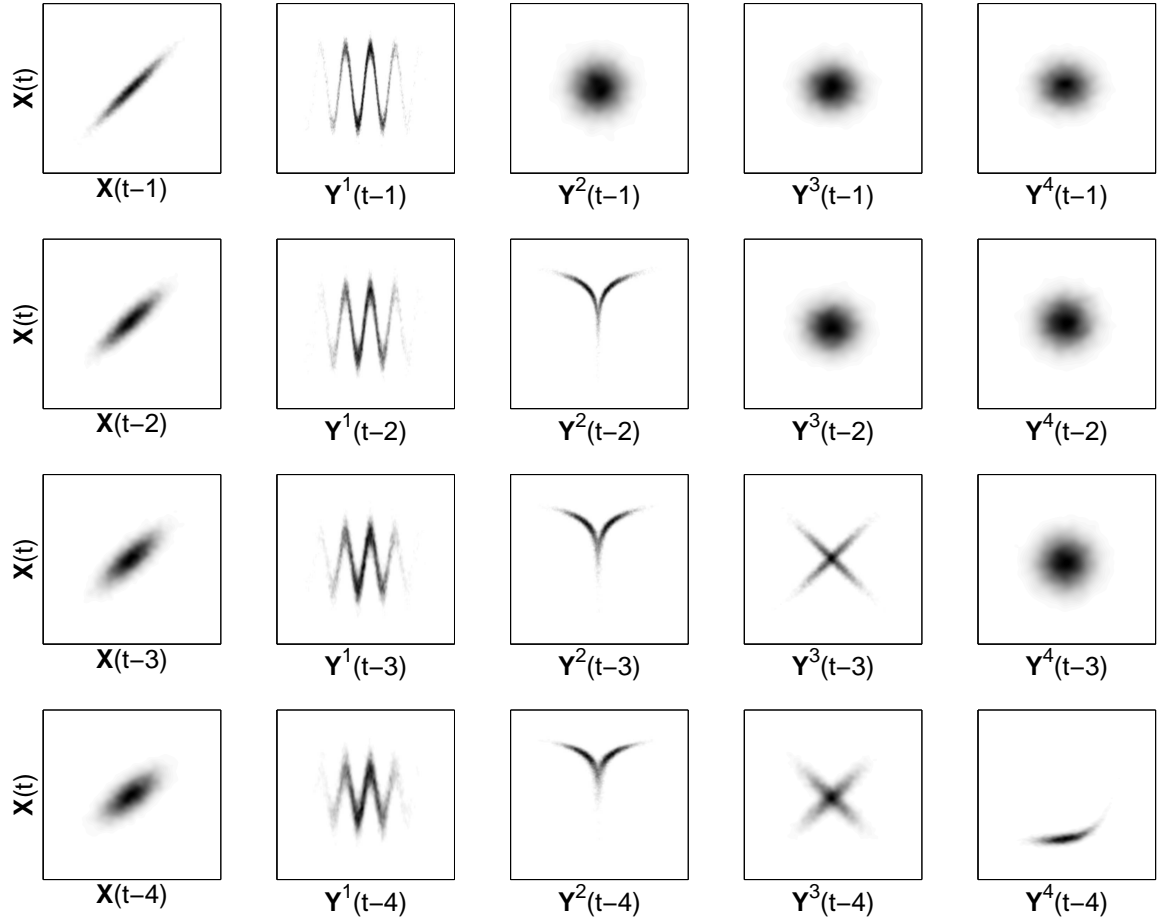


Figure 1: *Density plots of  $X_t$  against all lags of other variables. The system exhibits different types of non-linear and/or non-monotonous relationships. All other inter-variable relationships excluded from this figure display white-noise processes (see top right panels of this figure).*

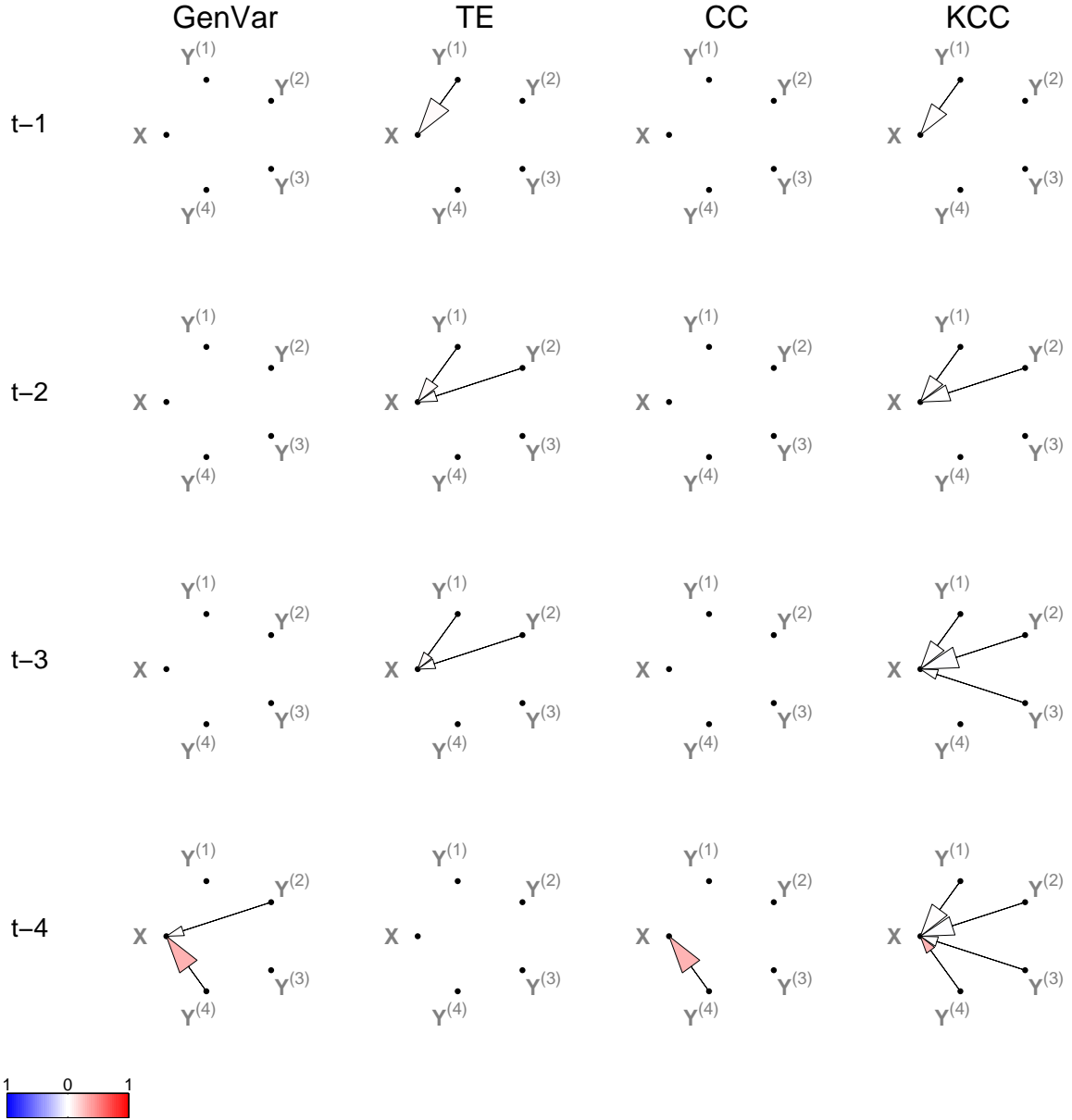


Figure 2: Results from the simulations in 3.1 displayed as networks where a critical value of  $\alpha = 0.001$  has been set to rule out non-causality. In the simulations, KCC outperforms the other methods as it successfully detects the planted causal signals. Both GenVar (generalized variance) and CC detect the monotonous exponential signal. Transfer entropy (TE) detects the causal signals from  $Y^{(1)}$  and  $Y^{(2)}$  to  $X$  in up to  $k = 3$  lags but fails to discover the remaining signals in higher dimensions. The size of the arrows displays the relative distance of the empirical causal measure from the 99th percentile of the distributions under the null hypotheses. The color of the arrow displays an aggregate score of rank correlations (the rank correlation of average ranks of the two vectors).

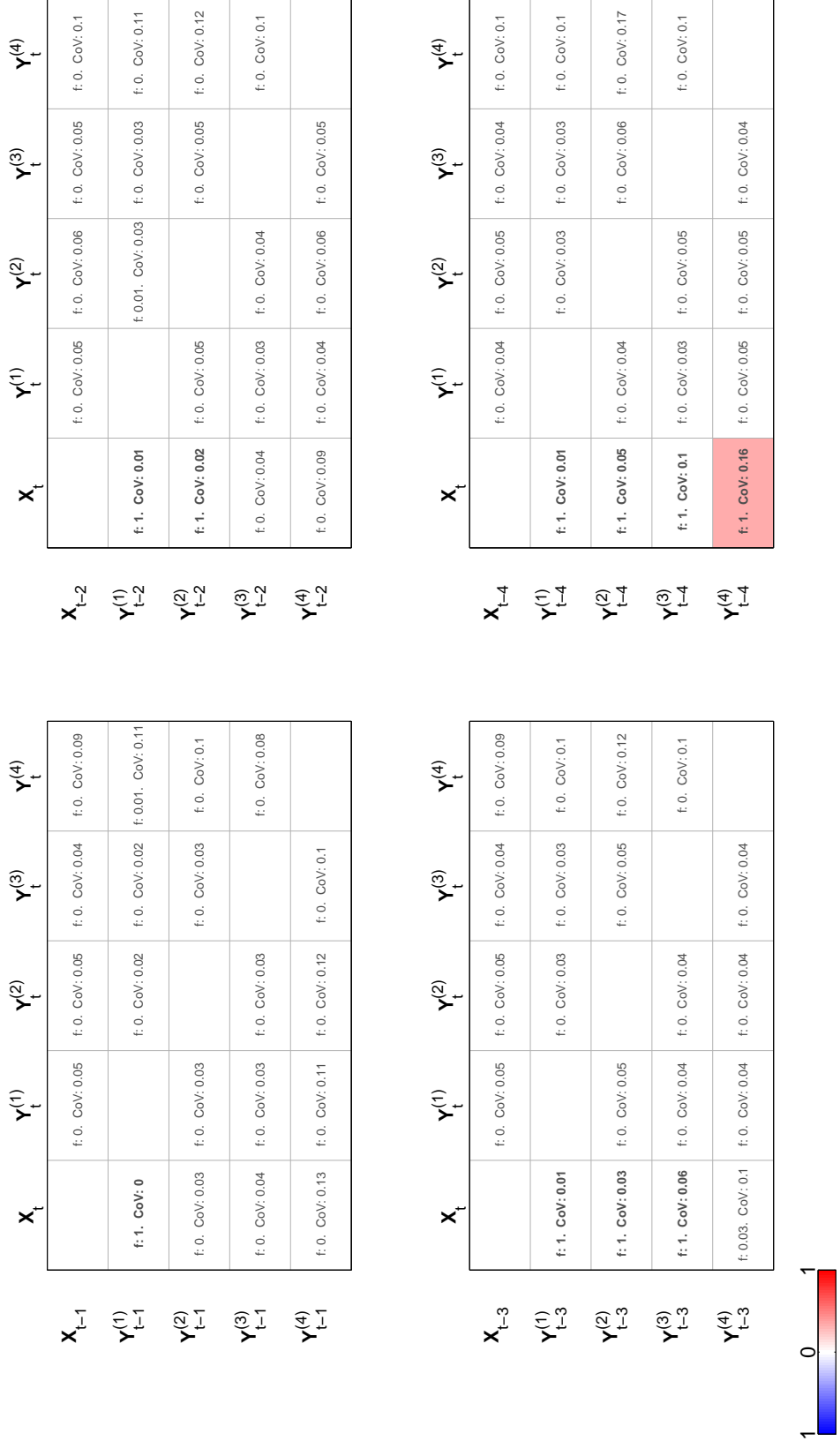


Figure 3: Results on the performance of KCC on 100 realizations of the system presented in 3.1. Causal signals go from row to column,  $f$  denotes the discovery rate of a causal signal, and  $\text{CoV}$  denotes the coefficient of variation of the KCC scores. The color-coding is based on aggregate score of rank correlations (the rank correlation of average ranks of the two vectors).

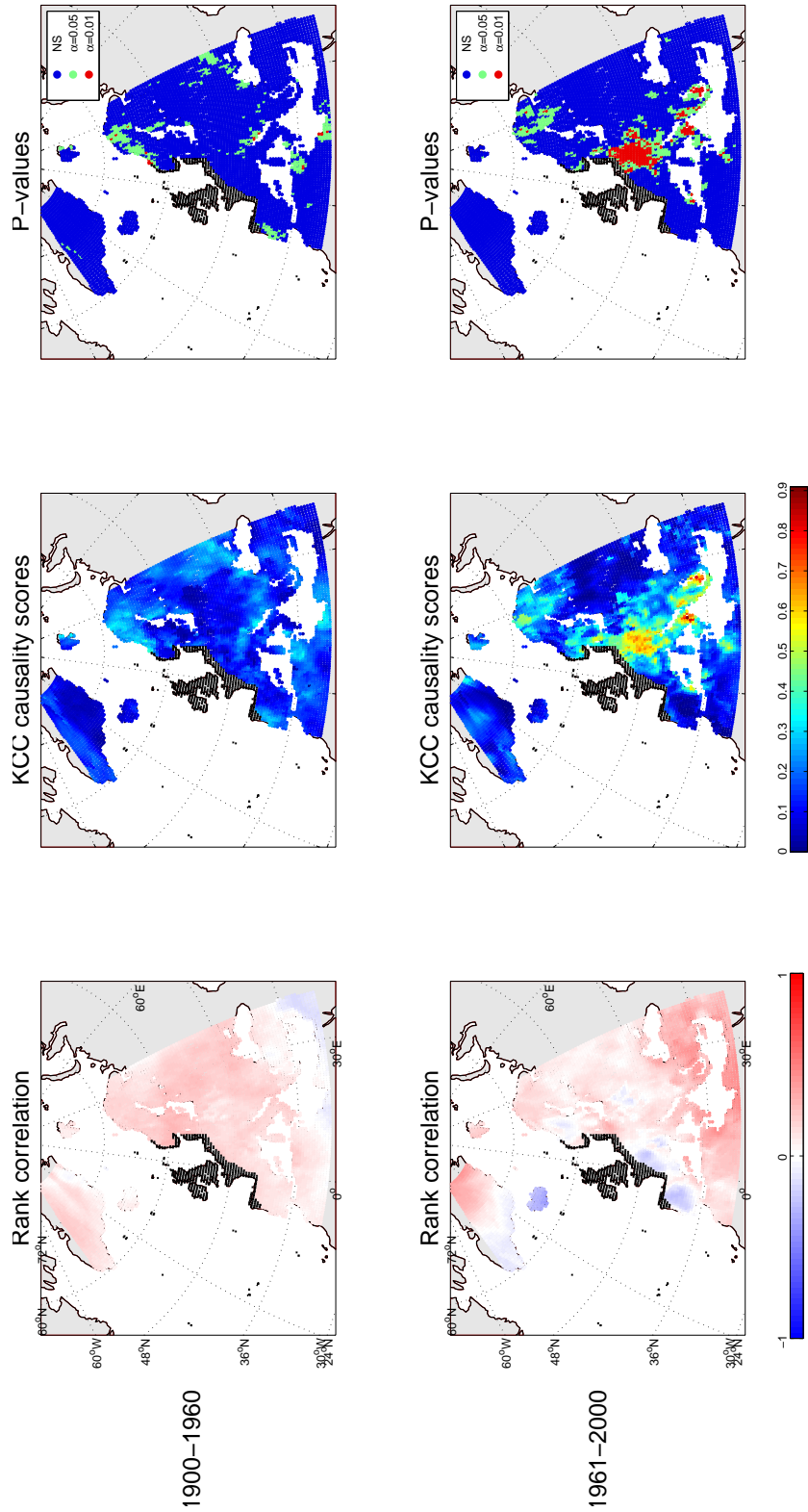


Figure 4: Analysis of Wiener-Granger causality based on the hypothesis in (21). The black-crossed is the North Atlantic region. The rank correlations present the correlations between temperature and precipitation in the North Atlantic region in March, and the temperature in rest of the continent in April. P-values are evaluated using permutation resamplings under the null hypothesis. Maps from the *M\_Map* toolbox by Rich Pawlowicz.