<http://www.diva-portal.org>

This is the published version of a paper published in *Applied Mathematical Sciences*.

Citation for the original published paper (version of record):

Jafari-Mamaghani, M. (2013)

Non-parametric analysis of Granger causality using local measures of divergence.

Applied Mathematical Sciences, 7(83): 4107-4236

<http://dx.doi.org/10.12988/ams.2013.35275>

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-101588>

Non-parametric Analysis of Granger Causality Using Local Measures of Divergence

Mehrdad Jafari-Mamaghani

Department of Mathematics, Stockholm University, Stockholm, Sweden
Department of Biosciences & Nutrition, Karolinska Institutet, Huddinge, Sweden

Copyright © 2013 Mehrdad Jafari-Mamaghani. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

The employment of Granger causality analysis on temporal data is now a standard routine in many scientific disciplines. Since its inception, Granger causality has been modeled using a wide variety of analytical frameworks of which, linear models and derivations thereof have been the dominant choice. Nevertheless, a body of research on Granger causality and its applications has focused on non-linear and non-parametric models. One common choice for such models is based on employment of multivariate density estimators and measures of divergence. However, these models are subject to a number of estimations and tuning components that have a great impact on the final outcome. Here we focus on one such general model and improve a number of its tuning bodies. Crucially, we *i)* investigate the bandwidth selection issue in kernel density estimation, and *ii)* discuss and propose a solution to the sensitivity of estimated information theoretic measures of divergence to non-linear correspondence. The resulting framework of analysis is evaluated using varied series of simulations.

Keywords: Granger causality, Kernel density estimation, Bandwidth selection, Information theory, Jensen-Shannon divergence, Mutual information

1 Introduction

Analysis of causality, as a natural extension of correlative analysis, has been of great interest in many scientific disciplines. Causal dynamics can be quantified

and studied in as diverse areas as social events behind rising levels of criminality, the effects of protein expression on cellular motility, the causes behind growing concentrations of carbon dioxide in the atmosphere, the driving forces leading to increased inflation, and the effects of drug treatments on neuronal activity.

The general concept of causality has a history stretching as far back as that of philosophical thinking. However, quantitative concepts of causality are relatively new and have been subject to much debate. Given the wide variety of views on quantitative definitions of causality, it is of little surprise that different schools of thought on this matter have generated a considerable spectrum of quantitative frameworks to define, model and analyze data-driven causal phenomena. Among these, Bayesian networks, differential equation-driven systems analysis and Granger causality have been the dominant frameworks of causal modeling and analysis [17,28]. Here, we address the concept of Granger causality.

The concept of Granger causality was formulated in different lights by the works of [18,48] and consolidated by Clive W.J. Granger in [15]. Granger causality is a particular definition of causality where using temporal resolution, a variable is said to Granger-cause another if the earlier values of the former can enhance the prediction of the present value of the latter in the presence of the latter's earlier values. This particular definition of causality presumes a temporal signal asymmetry where the cause precedes the effect and where the information embedded in the causal variable about the occurrence of the effect conditioned on all other embedded information is unique [15,22]. Expressed using the mathematical language of probability theory, under H_0 , given k lags and variables A, B and C , $\{B\}$ does not Granger-cause $\{A\}$ at observation index t , if

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

The statement above can be tested by comparing the two conditional probability densities (CPDs) below [12]:

$$f(A_t | A_{t-1}, \dots, A_{t-k}, C_{t-1}, \dots, C_{t-k}) \quad (2)$$

$$f(A_t | A_{t-1}, \dots, A_{t-k}, B_{t-1}, \dots, B_{t-k}, C_{t-1}, \dots, C_{t-k}) \quad (3)$$

Under H_0 , where $\{B\}$ does not Granger-cause $\{A\}$, the CPDs (2) and (3) are identical. For the sake of convenience let us implement the following substitutions: $X = A_t, Y = \{B\}_{t-1}^{t-k}$ and $Z = \{\{A\}_{t-1}^{t-k}\{C\}_{t-1}^{t-k}\}$. Thus, it is understood that all formulations, estimations and tests, regardless of their appearance are easily identifiable with any arbitrary multivariate setting. The conditional density functions in (2) and (3) are thus expressed more economically as $f_{X|Z}$ and $f_{X|YZ}$, respectively. Accordingly, (1) can be redefined as:

$$H_0 : X \perp Y | Z \quad (4)$$

Since the introduction of the concept of Granger causality in 1969, for which Granger received the Nobel Memorial Prize in Economics in 2003, this particular definition of causality has led to a wide array of research focused on its theoretical formulation, data-driven analyses, and expansions of practical routines using various statistical tools. Models of Granger causality have extensively been reviewed in [22, 30, 36, 40]. Most prominently, Granger causality has had a long range of applications within the econometric discipline [11, 14, 15, 39, 43]. Another prominent field of application includes biologically oriented domains such as biological networks inference, and systems modeling and analysis [21, 32, 35, 46].

In this study we will enhance the estimates of multidimensional probability densities used in non-parametric analysis of Granger causality by improved means of bandwidth selection, and address the sensitivity of estimated information theoretic measures to non-linear relationships through the employment of local estimates. The arrangement of the study is as follows. In Methods, we present the elementary models of Granger causality, review the non-parametric extensions using multivariate density estimation and measures of divergence, and address the bandwidth selection and sensitivity issues mentioned above. In Results, we investigate the performance of the improved non-parametric framework using series of simulations. Lastly, in Discussion we draw upon the discussed themes and conclude the study.

2 Methods

2.1 Parametric models

Methods of linear regression have been the most frequently employed tools to model and test the presence of Granger causality. Using linear regression, the hypothesis in (1), omitting the variable(s) C for notational convenience, is tested using the models

$$H_0 : A_t = \alpha_0 + \sum_{i=1}^k \beta_i A_{t-i} + \epsilon_t$$

$$H_1 : A_t = \alpha_1 + \sum_{i=1}^k \beta_i A_{t-i} + \sum_{i=1}^k \gamma_i B_{t-i} + \eta_t$$

where α are the regression intercept coefficients, β and γ are the regression variable coefficients, and the residual terms ϵ_t and η_t are independent and identically distributed according to a standard Gaussian $N(0, \sigma^2)$. See [13] for a generalization of the models above.

Among other techniques, the models above can be tested using the Granger-

Sargent test [15], also known as the structural Chow test in econometric literature [9]:

$$F = \frac{(RSS_\epsilon - RSS_\eta)/k}{RSS_\eta/(n - 2k)} \sim F(k, n - 2k) \quad (5)$$

where RSS_ϵ is the 'restricted' residual sum of squares under H_0 , RSS_η is the 'unrestricted' residual sum of squares under H_1 , n is the number of observations, and k is the number of included lags.

Naturally, these models assume linear specifications of the functional form between the regressors and the response variable. Additionally, given the widespread usage of ordinary least squares' (OLS) regression, the models demand the fulfillment of the standard assumptions of homoscedasticity, error normality, and lack of multicollinearity. A routine practice to relax the latter assumption is achieved by principal components regression or partial least squares regression at the possible expense of predictive power. A partial circumvention of the other restrictions is outlined in the following. An approach to relax the strict linearity of the models above is to use functional derivations or higher-dimensional projections of variables. Thus, the hypothesis in (1) can be tested using the models

$$H_0 : A_t = \alpha'_0 + \sum_{i=1}^k \sum_{j=1}^{J_i} \beta_{ij} f_{ij}(A_{t-i}) + \epsilon'_t$$

$$H_1 : A_t = \alpha'_1 + \sum_{i=1}^k \sum_{j=1}^{J_i} \beta_{ij} f_{ij}(A_{t-i}) + \sum_{i=1}^k \sum_{j=1}^{J_i} \gamma_{ij} f_{ij}(B_{t-i}) + \eta'_t$$

where f_{ij} denotes the j th functional derivation of a variable at lag $t - i$, and J_i represents the number of such derivations for each variable at each lag $t - i$. Derivations and projections of this type can naturally be applied to the response variable as well, leading to multivariate models of regression. Typical derivations and projections include polynomial expansions, power transformations, spline-based expansions, and radial basis function expansions. The hypothesis modeled above can be similarly tested using the Granger-Sargent statistic in (5). Specific studies in the modeling and application of non-linear models of regression in the context of Granger causality include [1, 5, 7, 27].

Although offering an enhanced degree of flexibility, models of non-linear regression as formulated above differ from those of linear regression merely in the sense that they model the linear correspondences between derivations of the response and the regressors. Hence, the obtained specifications of the relationships will still be linear or piece-wise linear. Whether or not the modeled relationships are the true underlying patterns of correspondence is not a trivial task to verify. Moreover, higher-dimensional projections of data may need ade-

quate treatments as they often lead to overparametrization and ill-conditioned covariance matrices.

2.2 Non-parametric models

As discussed in the previous section, models of linear regression and the corresponding parametric tests lose their practical validity where assumptions are violated. As a simple example, regression via least squares' techniques with the Granger-Sargent test in (5) perform rather poorly in the presence of variables with skewed distributions. More important however, is the potential exacerbation of performance inadequacy in the case of linear model specification where the true forms of correspondence between the response and the regressors may be non-linear.

Based on these concerns, efforts to improve the modeling of Granger causality have focused on relaxing the assumptions of parametricity in general, and liberation from linear forms of correspondence in particular. Here we will focus on the circumvention of these assumptions by using means of kernel density estimation and non-parametric (information theoretic) measures of divergence. Frameworks of this breed have been reviewed and discussed to varying extents in [22, 36]. In the current study however, our aim is to improve some of the overlooked aspects of non-parametric and information theoretic-driven modeling of Granger causality. These include the practices involved in multivariate kernel density estimation and the sensitivity of information theoretic measures to non-linear relationships.

In the following we outline our framework to test the hypothesis formulated in (4) by comparing the CPDs (2) and (3). Consequently, the procedure of modeling and testing (4) can be divided into three domains: *i*) estimation of the CPDs, *ii*) choosing a measure of divergence and, *iii*) utilization of a suitable framework for tests of significance.

2.2.1 Estimating CPDs: Kernel density estimation

Traditionally, histogram density estimators based on equidistant, spline-based or adaptive partitioning of the observation space, have been the preferred solution for estimating probability density functions. Since their introduction however, kernel density estimation (KDE) has become the dominant solution to probability density estimation due to greater flexibility and superior performance [34, 38, 47]. More specifically, when compared to histogram density estimators, the KDE approach leads to better rates of convergence and less sensitivity to binning [41].

Having obtained independent and identically distributed samples $(x_1, \dots, x_n)$ from a distribution with an unknown density f , the estimation of f using KDE

reduces to the following:

$$\hat{f}_h(X) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X - x_i}{h}\right)$$

where $K(\cdot)$ is a kernel function and h is a bandwidth, a smoothing parameter. Common kernel functions are the uniform, triangular, Epanechnikov, or the Gaussian [20, 38]. The latter function, the standard Gaussian kernel $K(x) = \phi(x) = e^{-x^2/2}/\sqrt{2\pi}$ will be our choice throughout this study.

Thus, using KDE, $f_{X|Z}$ and $f_{X|YZ}$ are estimated according to the following:

$$\hat{f}_{X|Z} = \frac{\hat{f}_{XZ}}{\hat{f}_Z} = \frac{C_{XZ}^{-1} \left[\sum_{i=1}^n K\left(\frac{X - x_i}{h^{(X)}}\right) K\left(\frac{Z - z_i}{h^{(Z)}}\right) \right]}{C_Z^{-1} \left[\sum_{i=1}^n K\left(\frac{Z - z_i}{h^{(Z)}}\right) \right]} \quad (6)$$

$$\hat{f}_{X|YZ} = \frac{\hat{f}_{XYZ}}{\hat{f}_{YZ}} = \frac{C_{XYZ}^{-1} \left[\sum_{i=1}^n K\left(\frac{X - x_i}{h^{(X)}}\right) K\left(\frac{Y - y_i}{h^{(Y)}}\right) K\left(\frac{Z - z_i}{h^{(Z)}}\right) \right]}{C_{YZ}^{-1} \left[\sum_{i=1}^n K\left(\frac{Y - y_i}{h^{(Y)}}\right) K\left(\frac{Z - z_i}{h^{(Z)}}\right) \right]} \quad (7)$$

where $C_{\cdot}$ are the normalizing constants, i.e.

$$C_{XYZ} = N \left(\prod_{j=1}^{d_X} h_j^{(X)} \right) \left(\prod_{j=1}^{d_Y} h_j^{(Y)} \right) \left(\prod_{j=1}^{d_Z} h_j^{(Z)} \right) \quad (8)$$

where $N = n_{d_X} + n_{d_Y} + n_{d_Z}$ is the total number of observations used in the estimation, d_X is the number of dimensions of X , and $h^{(X)}$ are the bandwidths associated with the corresponding dimensions of X ; resulting in an individual bandwidth for every lag of every variable in the KDE of the CPDs. The other definitions follow accordingly.

Varieties of KDE in non-parametric modeling of Granger causality have been previously employed to test the hypothesis in (4) in a number of studies [8, 21, 35]. Specifically, the kernel function chosen in [8] is $K(x) = (3 - x^2) \phi(x)/2$ and the bandwidths are evaluated using a 'rule of thumb'. Notably, the choice of bandwidths is an issue that demands a discussion.

Bandwidth selection

Adequate bandwidth selection in kernel density estimation of multivariate distributions is critical to obtaining reliable estimations. The bandwidth, as

a smoothing parameter, determines the number of observations included in 'windows' in the process of KDE. Specification of the bandwidth size is a clear bias-variance tradeoff where relatively small bandwidths lead to high variance and low bias, and relatively large bandwidths to the opposite [20].

Currently, choosing the bandwidth parameter in a wide variety of applications is determined by a 'rule of thumb' or other application-specific evaluations. Unsurprisingly, such frameworks fail to generalize and when used without critical reflection can lead to erroneous and biased estimations. This issue has to an extent been addressed by other more adaptive methods designed to enhance the unbiasedness of KDE [19, 23, 45]. However, none of these approaches qualifies as a purely non-parametric practice yielding satisfactory density estimations regardless of underlying distributions.

Here, in the following employments of KDE, bandwidths are chosen using the plug-in bandwidth selection method proposed in [4] based on linear diffusion processes. Using simulated data, the bandwidths produced according to [4] return reliable and accurate results and consistently outperform the traditional bandwidth selection methods (see Figure 1).

The simulations in Figure 1 are based on two axes of variation: distribution bimodality and mode shape. The source distributions are two-component mixtures of Gaussian distributions with 1000 randomly generated numbers and equal mixing proportions. The bimodality of the distribution is determined by distancing the two Gaussian components and the mode shape is determined by altering the variance of one of the components. Both of these variations are performed progressively for 20·20 bins as seen in Figure 1. The obtained residuals represent the pointwise standardized deviations between the estimated densities and the underlying mixture model density. The fixed bandwidth of choice has been $(4\hat{\sigma}^5/3n)^{0.2}$, where $\hat{\sigma}$ is the standard deviation of the sample and n is the number of samples. This bandwidth is also known as Silverman's rule of thumb [38]. As seen in Figure 1, the adaptive bandwidths proposed by [4] clearly outperform the fixed bandwidths in terms of unbiasedness. Thus, in the remainder of this study all bandwidths in all instances of KDE are realized using the method in [4].

2.2.2 Measures of divergence

As stated in the Introduction, the existence of Granger causality is equivalent to an inequality between the conditional probability densities $f_{X|Z}$ and $f_{X|YZ}$. Under the null hypothesis (4) the CPDs above have to be equal. This plausible equality or its lack thereof, can be quantified using a variety of measures of divergence. Among others, the weighted Hellinger distance and the Euclidean distance have been used to quantify the divergence between the kernel density estimated CPDs in [8], of which the latter will be employed in a number of our

simulations. More specifically, the Euclidean distance between (7) and (6) is defined as [44]:

$$d_E = n^{-1} \left[\sum_{i,j} \|(f_{X|YZ})_i - (f_{X|Z})_j\| - \frac{1}{2} \sum_{i,j} \|(f_{X|YZ})_i - (f_{X|YZ})_j\| - \frac{1}{2} \sum_{i,j} \|(f_{X|Z})_i - (f_{X|Z})_j\| \right] \quad (9)$$

where $(f_{\cdot|\cdot})_i$ is the i th realization of $(f_{\cdot|\cdot})$, and n is the number of such realizations.

In the remainder the focus will be on measures of divergence derived from the field of information theory. This focus is motivated by the inherent non-parametricity of information theoretic measures of divergence. Firstly, given empirically estimated distribution functions, there is no need for distribution specificity. Secondly, information theoretic measures operate in an evidence-based manner allowing detection of non-linear relationships, the practical validity of which will be discussed shortly.

The first discussed measure for quantification of divergence between $f_{X|YZ}$ and $f_{X|Z}$ is the Jensen-Shannon divergence introduced in [26]. Given the right weighting parameters, the Jensen-Shannon divergence can be regarded as the symmetrized version of the Kullback-Leibler divergence [25, 26]. However, before a thorough elaboration of these measures, a review on some of the basic concepts in information theory may be useful.

In information theory the (differential) Shannon entropy of a random variable X with a continuous probability distribution f_X with support on $\mathcal{X}$ is defined as [10, 37]:

$$H(X) \equiv -E[\log_b f_X] = - \int_{\mathcal{X}} f_X \log_b f_X dx \quad (10)$$

where b is the base of the logarithm determining the terms in which the entropy is measured (e.g. $b = 2$ for bits and $b = e$ for nats). The higher the value of entropy the higher the uncertainty associated with the outcomes of the studied random variable. The most illustrative demonstrations are those of bernoulli trials such as coin tosses. A symmetric coin ($p_1 = p_2 = 0.5$) yields the highest entropy value whereas values of entropy decrease with increasing coin asymmetry; hence decreasing uncertainty.

The Kullback-Leibler divergence between two probability distribution functions f_X and g_X with support on $\mathcal{X}$ is defined as [10, 25]:

$$D_{KL}(f_X||g_X) \equiv \int_{\mathcal{X}} f_X \log_b \frac{f_X}{g_X} dx \quad (11)$$

Note that replacing the arguments f_X and g_X in (11) can lead to a different quantification of divergence. Therefore, the Kullback-Leibler divergence is not

regarded as a symmetric measure of divergence.

Given the definitions above, the Jensen-Shannon divergence between two probability distributions f_X and f_Y representing two stochastic variables X and Y , respectively, with identical support is defined by

$$D_{JS}(f_X||f_Y) \equiv \pi_X D_{KL}(f_X||m_{XY}) + \pi_Y D_{KL}(f_Y||m_{XY}) \quad (12)$$

where $m_{XY} = [f_X + f_Y]/2$, and π_i denotes the weights assigned to each distribution subject to $\pi_X + \pi_Y = 1$. In our applications, unless otherwise stated, $\pi_i = 1/d$ where d is the number of dimensions. Moreover, $0 \leq D_{JS}(\cdot||\cdot) \leq \log_b(b)$. Expressed in terms of Shannon entropy, (12) can be redefined as:

$$D_{JS}(f_X||f_Y) \equiv H(\pi_X X + \pi_Y Y) - \pi_X H(X) - \pi_Y H(Y) \quad (13)$$

Thus, with $\pi = 0.5$, the symmetrized divergence between the estimated CPDs (7) and (6) is evaluated according to:

$$\begin{aligned} D_{JS}(f_{X|YZ}||f_{X|Z}) &= H(\pi[X|YZ] + \pi[X|Z]) - \pi H(X|YZ) - \pi H(X|Z) \\ &= - \int_{\mathcal{X}\mathcal{Y}\mathcal{Z}} [\pi f_{X|YZ} + \pi f_{X|Z}] \log_b [\pi f_{X|YZ} + \pi f_{X|Z}] dx dy dz \\ &\quad + \pi \int_{\mathcal{X}\mathcal{Y}\mathcal{Z}} f_{X|YZ} \log_b [f_{X|YZ}] dx dy dz \\ &\quad + \pi \int_{\mathcal{X}\mathcal{Z}} f_{X|Z} \log_b [f_{X|Z}] dx dz \end{aligned} \quad (14)$$

Under $H_0 : X \perp Y|Z$, it is easily seen that $D_{JS}(\hat{f}_{X|YZ}||\hat{f}_{X|Z}) = 0$.

Another information theoretic measure of divergence frequently used in the context of Granger causality has been the conditional mutual information. In fact, there is an illuminating relationship between (14) and conditional mutual information which is elaborated in the Appendix 5.1.

The conditional mutual information of X and Y given Z , with supports on $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ respectively, can be expressed as:

$$\begin{aligned} I(X; Y|Z) &= H(X|Z) + H(Y|Z) - H(XY|Z) \\ &= H(XZ) + H(YZ) - H(XYZ) - H(Z) \\ &= H(X|Z) - H(X|YZ) \\ &= \int_{\mathcal{X}\mathcal{Y}\mathcal{Z}} f_{XYZ} \log_b \frac{f_{XYZ}}{f_{X|Z}f_{Y|Z}} dx dy dz \end{aligned} \quad (15)$$

The conditional mutual information is symmetric: $I(X; Y|Z) = I(Y; X|Z)$. Furthermore, $I(X; Y|Z) \geq 0$ with equality if and only if X and Y are independent. Lastly, $I(X; X|Z) = H(X|Z)$.

As the conditional entropy of a set of variables could vary, the conditional

mutual information in (15) should be normalized to compensate for possible fluctuations. In this study, we will employ the normalized conditional mutual information and conditional symmetric uncertainty:

$$NI(X; Y|Z) = \frac{I(X; Y|Z)}{\sqrt{H(X|Z)H(Y|Z)}} \quad (16)$$

$$SU(X; Y|Z) = \frac{2 \cdot I(X; Y|Z)}{H(X|Z) + H(Y|Z)} \quad (17)$$

Other studies outside of the context of Granger causality focused on normalization of mutual information include [6, 42, 49]. It should be noted that a number of information theoretic measures used in the context of Granger causality can be deducted to a derivation of conditional mutual information.

One of these measures is the widely used 'transfer entropy' as proposed in [33], which can be reparametrized as the conditional mutual information [22, 36]. Interestingly, the transfer entropy is shown to have a functional relationship with the linear estimators of Granger causality [3, 29]. Other measures coined under the term 'directed influence' also fall under this category as derivations of conditional mutual information [36].

2.2.3 Sensitivity to linearity

Based on our observations and as put forward in [24, 31], information theoretic measures estimated via KDE show a quantifiable and non-random sensitivity to non-linear patterns of correspondence. Thus, given the same degree of noise, information theoretic measures estimated via KDE assign lower scores to non-linear relationships than those of a linear manner. As this contradicts the non-parametric nature of information theoretic measures a series of different approaches have been devised to circumvent this issue.

As outlined in [31], limiting the sensitivity of information theoretic measures (e.g. mutual information) to non-linearity may be achieved by means of domain partitioning. However, the partitioning devised in [31] demands a large number of observations and due to its extensive partitioning is limited to the bivariate case. It is understood that in the present case of time-series modeling using two or more variables and one or more lags, the smallest number of dimensions available for partitioning is three. Moreover, as the number of variables and lags increase, the 'curse of dimensionality' will swiftly necessitate the collection of ever larger deposits of observations.

To circumvent this issue, we propose local estimates of the measures above

(14), (16) and (17) as follows:

$$D_{JS}^{\mathcal{N}}(f_{X|YZ}||f_{X|Z}) = \frac{1}{n} \sum_{i=1}^n D_{JS} \left(f_{X^{(\mathcal{N}_i)}|Y^{(\mathcal{N}_i)}Z^{(\mathcal{N}_i)}} || f_{X^{(\mathcal{N}_i)}|Z^{(\mathcal{N}_i)}} \right) \quad (18)$$

$$NI^{\mathcal{N}}(X; Y|Z) = \frac{1}{n} \sum_{i=1}^n NI \left(X^{(\mathcal{N}_i)}; Y^{(\mathcal{N}_i)}|Z^{(\mathcal{N}_i)} \right) \quad (19)$$

$$SU^{\mathcal{N}}(X; Y|Z) = \frac{1}{n} \sum_{i=1}^n SU \left(X^{(\mathcal{N}_i)}; Y^{(\mathcal{N}_i)}|Z^{(\mathcal{N}_i)} \right) \quad (20)$$

where n is the number of observations, $\mathcal{N}_i$ denotes the neighborhood of observation i and vectors $X^{(\mathcal{N}_i)}$, $Y^{(\mathcal{N}_i)}$ and $Z^{(\mathcal{N}_i)}$ represent the observation values in neighborhood $\mathcal{N}_i$. Naturally, the ability of these estimates to capture non-linear relationships depends on the size of the neighborhood. Adequate neighborhood sizes should be chosen subject to their ability to capture local relationships. Small neighborhood sizes would potentially suffer from large variance whereas neighborhoods of large proportions might be increasingly biased and fail to capture local structures. When applied to time-series modeling, neighborhood selection is most intuitively determined by basing it on the domain of the response variable (the 'effect' of the 'cause' variables). Simulations (see Figure 8) confirm the ability of these estimates to detect non-linear relationships.

2.2.4 Tests of significance

Although there are distribution parametrizations for some information theoretic measures under specific conditions [36], given our aim to constitute a non-parametric framework for modeling and testing, we choose to employ bootstrap resampling to create estimations of probability distribution for the chosen measures of divergence under the null hypotheses. The bootstrap resampling is conducted B times with replacement under $H_0 : X \perp Y|Z$ where at each instance of resampling the empirical observations in Y are randomly permuted to artificially create the independence stated under H_0 . Comparing the empirically quantified measure of interest with the bootstrapped distribution under H_0 yields the relevant p-values. Naturally, depending on the chosen measure, violations of H_0 can occur under the lower, e.g. $D_{JS}(f_{X|YZ}||f_{X|Z})$, or above the upper quantiles, e.g. $I(X; Y|Z)$, of the bootstrapped distributions.

3 Results

The following series of simulations are designed to evaluate the performance of the hitherto discussed framework of Granger causality analysis. In the

bivariate and functional time-series, the parametric framework stated in the Methods has also been included. The focus in the functional time-series is on (19) as this measure performs nearly identically as (18) and (20). In the multivariate time-series, the aim is to investigate the ability of the framework in detection of causal links in high-dimensional spaces.

3.1 Bivariate time-series

In the simulated bivariate time-series only one variable is designed to be self-generative. More specifically, we define two variables X and Y , and denote the temporal index by t , $x_t = a$ and $y_t = b$ where a and b are random numbers drawn from the standard Gaussian distribution. The remaining is defined as:

$$\begin{cases} x_{t-i} = a + i \cdot \epsilon & i = 1, \dots, k \\ y_{t-i} = a + g_c(i) \cdot \epsilon & i = 1, \dots, k \end{cases}$$

where $\epsilon \sim N(0, 1)$, $k = 10$ and $g_c(i) = c \cdot i$ determines the degree to which lags of y are designed to correlate with x_t . Common choices for c in the following simulations include 0.1, 1, 2, 5 and 10. Each realization of X and Y consists of 100 $\{x_t\}$ and $\{y_t\}$, respectively. That is, 1100 sample points in X and Y . At each lag k the following hypotheses are tested:

$$\begin{aligned} H_0^X &: \{X_t\} \perp \{Y_{t-1}, \dots, Y_{t-k}\} | \{X_{t-1}, \dots, X_{t-k}\} \\ H_0^Y &: \{Y_t\} \perp \{X_{t-1}, \dots, X_{t-k}\} | \{Y_{t-1}, \dots, Y_{t-k}\} \end{aligned}$$

The hypotheses H_0^X and H_0^Y are tested via the classical linear regression framework in 2.1. In addition, using density estimation via KDE as devised in (7) and (6), the hypotheses above are tested by employing the Euclidean distance (9), the normalized conditional mutual information (16), the conditional symmetric uncertainty (17), the Jensen-Shannon divergence (12). The results based on 100 realizations of X and Y using $B = 1000$ in the bootstrap resamplings, are displayed in Figures 2, 3, 4, 5, 6, and summarized in Figure 7.

The arrangement of panels in Figures 2-6 is as follows: panels along the horizontal axis represent the noise levels 0.1, 1, 2, 5 and 10 whereas panels along the vertical axis represent the number of included lags: 1, 5 and 10. The gray probability masses represent the reference distributions under the null hypotheses H_0^X . These distribution are either evaluated analytically (as for the Granger-Sargent test), or obtained computationally using the bootstrap resampling scheme outlined above. The crosses on the x -axes of these probability masses represent the 'empirical' scores of the simulations. It is easily seen that increasing noise levels, regardless of the number of included lags, lead to less frequent rejections of the null hypotheses. That is, the detectability of the synthetic causal link between the two variables is degraded as a function

of added noise.

The summary of the results above under H_0^X and additionally under H_0^Y are represented in Figure 7 in terms of p -values for each noise level at each lag. Here, it is easily seen that the Euclidean distance d_E is outperformed by the Granger-Sargent test and the divergence measures under H_0^X . Additionally, the information theoretic measures of divergence outperform the Granger-Sargent test in the detection of causal relationships. Among the information theoretic measures, the Jensen-Shannon divergence performs most optimally.

3.2 Functional time-series

Here we define two variables X and Y , and denote the temporal index by t , $x_t = a$ and $y_t = b$ where a is a random number generated according to a $U(-2, 2)$ distribution and b is a random number drawn from the standard Gaussian distribution. The remaining is defined as:

$$\begin{cases} x_{t-i} = a + g_c(i) \cdot \epsilon & i = 1, \dots, k \\ y_{t-i} = \sin(a) + g_c(i) \cdot \epsilon & i = 1, \dots, k \end{cases}$$

where $\epsilon \sim N(0, 1)$, $k = 5$ and $g_c(i) = c \cdot i$ determines the degree to which lags of x and y are designed to correlate with x_t . In this series c is set to three different values: 1, 2.5 and 5. Each realization of X and Y consists of 50 $\{x_t\}$ and $\{y_t\}$, respectively. The null hypothesis is identical to that of the bivariate time-series, tested using the Granger-Sargent test and the neighborhood-based measure of divergence is normalized conditional mutual information as defined in (19) where the neighborhood size is chosen to include the 10 nearest neighbors to every observation. The results based on 100 simulations of X and Y using $B = 500$ in the bootstrap resamplings, are displayed in Figure 8.

The groups of scatterplots on the left side of Figure 8 represent the bivariate distributions of X at lag 0 against X and Y at lags $k = 1, \dots, 5$. The scatterplots are arranged vertically top down with increasing noise levels. The three panels on the right side in Figure 8 represent the yielded p -values from the simulations. Regardless of the noise level, the Granger-Sargent test fails to capture the functional causal link between X and Y whereas the neighborhood-based normalized conditional mutual information NI^N detects the relationship at all noise levels.

3.3 Multivariate time-series

The simulated multivariate time-series consists of four variables (nodes) W, X, Y , and Z and resembles a type of reverse design in its construction compared to the former bivariate cases.

More specifically, $w_t = a$, $x_t = b$, $y_t = c$ and $z_t = d$, and where a, b, c and

d are generated randomly according to the standard Gaussian distribution. Additionally:

$$\begin{cases} w_{t-k} = w_t + g_3(1) \cdot \epsilon \\ x_{t-k} = w_t + g_1(1) \cdot \epsilon \\ y_{t-k} = x_t + g_{1.5}(1) \cdot \epsilon \\ z_{t-k} = x_t + g_{0.15}(1) \cdot \epsilon + y_t + g_{0.3}(1) \cdot \epsilon \end{cases}$$

where $k = 10$ and where similarly, each ϵ is generated randomly according to a $N(0, 1)$. The remaining is constructed according to:

$$\begin{cases} w_{t-i} = w_{t-i-1} + g_3(k-i) \cdot \epsilon & i = 1, \dots, k-1 \\ x_{t-i} = w_{t-i-1} + g_1(k-i) \cdot \epsilon & i = 1, \dots, k-1 \\ y_{t-i} = x_{t-i-1} + g_{1.5}(k-i) \cdot \epsilon & i = 1, \dots, k-1 \\ z_{t-i} = z_{t-i-1} + g_{0.5}(k-i) \cdot \epsilon & i = 1, \dots, k-1 \end{cases}$$

This specific setting is designed to test whether the employed framework of KDE, divergence quantification and bootstrap tests of significance can capture time-resolved correlations in high dimensions. The cross-temporal correlations increase as more lags from the 'past' are included in the model. Consequently, during this process the space in which the data is embedded is inflated as more dimensions (lags) are added to the KDE.

The results of the simulation using $B = 1000$ bootstrap resamplings presented in Figure 9 and 10, based on the Jensen-Shannon divergence and normalized conditional mutual information respectively, will help to illuminate the motivation behind the specific design.

The demonstration of the simulation results in Figures 9 and 10 is arranged according to the following. The four nodes and their conditional causal links are represented for each progressive inclusion of lags. Any presence of a conditional causal link between any two nodes is marked by a directed arrow (for p -values < 0.05). Absences of conditional causal links are marked by dashed lines. The color-coding of the dashed lines denotes the rank correlation coefficients between one node at lag 0 and the other node at the most recently included lag. The color-coding of the arrows denotes the same quantity between the cause at the most recently included lag and the effect at lag 0.

As evident in Figures 9 and 10, the two frameworks perform equally well in unveiling the conditional causal relationships between the nodes. One slight difference however, is the order in which the causal links are detected. Closer investigations revealed that this phenomenon had its roots in the between-simulation differences in random number generation. As the causal signals have been embedded in 'late' lags, easing their detectability by looking 'farther back' in time, one significant outcome of these simulations is the power of information theoretic measures to detect conditional causal links in relatively high-dimensional spaces.

4 Discussion

Given the abundance of multivariate time-series data in social and life sciences, there is an evident and inherent interest in moving beyond stationary correlative analysis to dynamic analysis of time-resolved causal relationships. The concept of Granger causality, although not synonymous with causality itself, offers a powerful framework for analysis of causality in time-series data.

Regarding the elaborated methods to model Granger causality, techniques based on assumptions of parametricity (e.g. ordinary linear regression) are superior to non-parametric designs due to their lower computational demands and more accessible interpretability. Nevertheless, meeting the assumptions of parametric models may not always be feasible.

The framework outlined here in this study based on kernel density estimation (KDE) via adaptive bandwidths, information-theoretic measures of divergence, and bootstrap tests of significance, constitutes a fully non-parametric platform for analysis of Granger causality. Additionally, the inferior sensitivity of information theoretic measures derived via KDE to non-linear relationships has been solved using local neighborhood-based estimates of information theoretic measures.

Furthermore, the results from our extensive simulations, based on synthetic linear and non-linear (functional) causal relationships, confirm the ability of the discussed platform in detecting causal relationships subject to a varying array of signal to noise ratios. In conclusion, both frameworks succeed to reveal the underlying conditional causal relationships between the nodes, matching the engineering of the simulations.

On prospective directions of research, unbiased estimates of differential information theoretic measures of divergence should be assigned top priority. As in earlier studies, we have here seen that differential information theoretic measures of divergence estimated via KDE fail to adhere to their non-parametric premise in the presence of non-linear relationships. Possible solutions to this problem are discretization of the data space, altered kernel functions in the density estimates, or higher-dimensional projections of the data space combined with supervised regularizations.

Regardless of possible prospective routes, any relevance of further improvements should be judged upon the specific aim and question of the analysis.

Overall, we have shown further advances in non-parametric analysis of Granger causality, allowing such type of temporal data analysis to better take advantage of available information given the abundance of non-parametric data in numerous scientific fields such as econometrics, biology and climatology. The choice of non-parametric analysis of Granger causality is further motivated by the ever growing computational power, facilitating considerable increase in the efficiency of such frameworks of analysis.

Acknowledgements. The author wishes to thank Joanna Tyrcha and John G. Lock for insightful discussions and feedback. The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement # 258068; EU-FP7-Systems Microscopy NoE; from the Swedish Research Council grant # 340-2012-6011; and from the Center for Biosciences at Karolinska Institutet.

5 Appendix

5.1 Jensen-Shannon divergence and conditional mutual information

Suppose $X|Z \sim m_{XYZ}$ where $m_{XYZ} = [f_{X|YZ} + f_{X|Z}] / 2$ is a mixture distribution. Furthermore, regard Y as a binary variable where $Y = 0$ if $X|Z \sim f_{X|Z}$ and $Y = 1$ if $X|Z \sim f_{X|YZ}$ and $\pi = 0.5$. Then $D_{JS}(\hat{f}_{X|YZ} || \hat{f}_{X|Z})$ as defined in (12) can be derived further as:

$$\begin{aligned}
 D_{JS}(\hat{f}_{X|YZ} || \hat{f}_{X|Z}) &= \pi D_{KL}(f_{X|YZ} || m_{XYZ}) + \pi D_{KL}(f_{X|Z} || m_{XYZ}) \\
 &= \pi \int_{\mathcal{X}\mathcal{Y}\mathcal{Z}} f_{X|YZ} \log_b \left[\frac{f_{X|YZ}}{m_{XYZ}} \right] dx dy dz + \pi \int_{\mathcal{X}\mathcal{Y}\mathcal{Z}} f_{X|Z} \log_b \left[\frac{f_{X|Z}}{m_{XYZ}} \right] dx dy dz \\
 &= \pi \int_{\mathcal{X}\mathcal{Y}\mathcal{Z}} f_{X|YZ} [\log_b f_{X|YZ} - \log_b m_{XYZ}] dx dy dz \\
 &\quad + \pi \int_{\mathcal{X}\mathcal{Y}\mathcal{Z}} f_{X|Z} [\log_b f_{X|Z} - \log_b m_{XYZ}] dx dy dz \\
 &= -\pi \int_{\mathcal{X}\mathcal{Y}\mathcal{Z}} f_{X|YZ} \log_b m_{XYZ} dx dy dz - \pi \int_{\mathcal{X}\mathcal{Z}} f_{X|Z} \log_b m_{XYZ} dx dy dz \\
 &\quad + \pi \int_{\mathcal{X}\mathcal{Y}\mathcal{Z}} f_{X|YZ} [\log_b f_{X|YZ}] dx dy dz + \int_{\mathcal{X}\mathcal{Z}} f_{X|Z} [\log_b f_{X|Z}] dx dy dz \\
 &= -\pi \int_{\mathcal{X}\mathcal{Y}\mathcal{Z}} m_{XYZ} \log_b m_{XYZ} dx dy dz \\
 &\quad + \pi \int_{\mathcal{X}\mathcal{Y}\mathcal{Z}} f_{X|YZ} [\log_b f_{X|YZ}] dx dy dz + \int_{\mathcal{X}\mathcal{Z}} f_{X|Z} [\log_b f_{X|Z}] dx dy dz \\
 &= H(X|Z) - H(X|YZ) = I(X; Y|Z)
 \end{aligned}$$

where $I(X; Y|Z)$ is referred to as the conditional mutual information of X and Y given Z [10]. Similar proofs without using conditional densities can be found in [2, 16, 49].

References

- [1] 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.
- [2] Ziv Bar-Yossef. Sampling lower bounds via information theory. In *STOC '03: Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 335–344. ACM Press, 2003.
- [3] 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.
- [4] ZI Botev, JF Grotowski, and DP Kroese. Kernel density estimation via diffusion. *The Annals of Statistics*, 38(5):2916–2957, 2010.
- [5] David S Broomhead and David Lowe. Radial basis functions, multi-variable functional interpolation and adaptive networks. Technical report, DTIC Document, 1988.
- [6] Nathan D. Cahill. Normalized measures of mutual information with general definitions of entropy for multimodal image registration. In *Proceedings of the 4th international conference on Biomedical image registration, WBIR'10*, pages 258–268, Berlin, Heidelberg, 2010. Springer-Verlag.
- [7] Y. Chen, G. Rangarajan, J. Feng, and M. Ding. Analyzing multiple nonlinear time series with extended Granger causality. *Phys. Lett. A*, 324:26–35, 2004.
- [8] Nadine Chlaß and Alessio Moneta. Can graphical causal inference be extended to nonlinear settings? In Mauricio Suárez, Mauro Dorato, and Miklós Rédei, editors, *EPSA Epistemology and Methodology of Science*, pages 63–72. Springer Netherlands, 2009.
- [9] Gregory C. Chow. Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica*, 28(3):591–605, 1960.
- [10] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory 2nd Edition*. Wiley-Interscience, 2 edition, July 2006.
- [11] Michael Eichler. Granger causality and path diagrams for multivariate time series. *Journal of Econometrics*, 137(2):334–353, April 2007.
- [12] Jean-Pierre Florens and M Mouchart. A note on noncausality. *Econometrica: Journal of the Econometric Society*, pages 583–591, 1982.

- [13] John Geweke. Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association*, 77(378):304, 1982.
- [14] John Geweke. Inference and causality in economic time series models. In Z. Griliches and M. D. Intriligator, editors, *Handbook of Econometrics*, volume 2 of *Handbook of Econometrics*, chapter 19, pages 1101–1144. Elsevier, September 1984.
- [15] Clive W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–38, 1969.
- [16] Ivo Grosse, Pedro B. Galván, Pedro Carpena, Ramón R. Roldán, Jose Oliver, and H. Eugene Stanley. Analysis of symbolic sequences using the Jensen-Shannon divergence. *Physical Review E*, 65(4):041905+, March 2002.
- [17] Shuixia Guo, Christophe Ladroue, and Jianfeng Feng. Granger causality: Theory and applications. In Jianfeng Feng, Wenjiang Fu, and Fengzhu Sun, editors, *Frontiers in Computational and Systems Biology*, volume 15 of *Computational Biology*, pages 83–111. Springer London, 2010.
- [18] Trygve Haavelmo. The probability approach in econometrics. *Econometrica: Journal of the Econometric Society*, 1944.
- [19] Peter Hall, Tien Chung Hu, and James Stephen Marron. Improved variable window kernel estimates of probability densities. *The Annals of Statistics*, 23(1):1–10, 1995.
- [20] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, corrected edition, August 2009.
- [21] H. Hinrichs, H.J. Heinze, and M.A. Schoenfeld. Causal visual interactions as revealed by an information theoretic measure and fmri. *NeuroImage*, 31(3):1051–1060, 2006.
- [22] 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.
- [23] M. Jones, I. McKay, and T. Hu. Variable location and scale kernel density estimation. *Annals of the Institute of Statistical Mathematics*, 46(3):521–535, 1994.

- [24] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review. E*, 69(6 Pt 2), June 2004.
- [25] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- [26] J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, January 1991.
- [27] Daniele Marinazzo, Mario Pellicoro, and Sebastiano Stramaglia. Nonlinear parametric model for granger causality of time series. *Phys. Rev. E*, 73:066216, Jun 2006.
- [28] Judea Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, New York, NY, USA, 2000.
- [29] Mario Pellicoro and Sebastiano Stramaglia. Granger causality and the inverse ising problem. *Physica A*, 389(21):4747–4754, 2010.
- [30] Florin Popescu. Robust statistics for describing causality in multivariate time series. *Journal of Machine Learning Research - Proceedings Track*, pages 30–64, 2011.
- [31] David N. Reshef, Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, and Pardis C. Sabeti. Detecting Novel Associations in Large Data Sets. *Science*, 334(6062):1518–1524, dec 2011.
- [32] R. Salvador, A. Martínez, E. Pomarol-Clotet, S. Sarró, J. Suckling, and E. Bullmore. Frequency based mutual information measures between clusters of brain regions in functional magnetic resonance imaging. *NeuroImage*, 35(1):83–88, March 2007.
- [33] Thomas Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85:461–464, Jul 2000.
- [34] David W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization (Wiley Series in Probability and Statistics)*. Wiley, 1 edition, September 1992.
- [35] A.-K. Seghouane. Quantifying information flow in fmri using the kullback-leibler divergence. In *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, pages 1569–1572, 30 2011-april 2 2011.
- [36] 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.

- [37] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, July 1948.
- [38] B. W. Silverman. *Density Estimation for Statistics and Data Analysis (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, 1 edition, April 1986.
- [39] Christopher A. Sims. Money, Income, and Causality. *Amer Econ Rev*, 62(4):540–552, 1972.
- [40] Klaas E. Stephan and Alard Roebroeck. A short history of causal modeling of fMRI data. *NeuroImage*, January 2012.
- [41] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics (Oxford, England)*, 18 Suppl 2(suppl 2):S231–S240, October 2002.
- [42] Alexander Strehl and Joydeep Ghosh. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3:583–617, March 2003.
- [43] Liangjun Su and Halbert White. A nonparametric hellinger metric test for conditional independence. *Econometric Theory*, 24(04):829–864, August 2008.
- [44] Gábor J Székely and Maria L Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5, 2004.
- [45] George R. Terrell and David W. Scott. Variable Kernel Density Estimation. *The Annals of Statistics*, 20(3):1236–1265, 1992.
- [46] P. F. Verdes. Assessing causality from multivariate time series. *Phys. Rev. E*, 72:026222, Aug 2005.
- [47] M. P. Wand and M. C. Jones. *Kernel Smoothing (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, 1 edition, December 1994.
- [48] N. Wiener and P. Masani. The prediction theory of multivariate stochastic processes. *Acta mathematica*, 98:111–150, 1957.
- [49] Y. Y. Yao. Information-theoretic measures for knowledge discovery and data mining. In Karmeshu, editor, *Entropy Measures, Maximum Entropy Principle and Emerging Applications*, pages 115–136, Berlin, 2003. Springer.

Figure 2: **The Granger-Sargent test.** Results based on simulations of bivariate time-series for $k = 1..10$ lags and 5 levels of noise as defined in **Bivariate time-series**. The + signs on the x-axes represent the yielded scores from the Granger-Sargent test as formulated in (5). The gray probability mass is that of the null hypothesis H_0^X under simulated data using the same test statistic.

Figure 3: **The Euclidean distance.** Results based on simulations of bivariate time-series for $k = 1..10$ lags and 5 levels of noise as defined in **Bivariate time-series**. The + signs on the x-axes represent the yielded scores from the Euclidean distance as formulated in (9). The gray probability mass is that of the null hypothesis H_0^X under simulated data using the same measure of divergence.

Figure 4: **Normalized conditional mutual information.** Results based on simulations of bivariate time-series for $k = 1..10$ lags and 5 levels of noise as defined in **Bivariate time-series**. The + signs on the x-axes represent the yielded scores from the normalized conditional mutual information as formulated in (16). The gray probability mass is that of the null hypothesis H_0^X under simulated data using the same measure of divergence.

Figure 5: **Symmetric uncertainty.** Results based on simulations of bivariate time-series for $k = 1..10$ lags and 5 levels of noise as defined in **Bivariate time-series**. The + signs on the x-axes represent the yielded scores from the normalized conditional mutual information as formulated in (17). The gray probability mass is that of the null hypothesis H_0^X under simulated data using the same measure of divergence.

Figure 6: **The Jensen-Shannon divergence.** Results based on simulations of bivariate time-series for $k = 1..10$ lags and 5 levels of noise as defined in *Bivariate time-series*. The + signs on the x-axes represent the yielded scores from the Jensen-Shannon divergence as formulated in (13). The gray probability mass is that of the null hypothesis H_0^X under simulated data using the same measure of divergence.

Figure 7: **Summary of the bivariate simulations.** The aggregate results of the simulated bivariate time-series as defined in **Bivariate time-series** demonstrated using p-values under the null hypotheses. The five horizontal panels represent the five different noise levels (c) implemented in the simulations. The upper panels represents the results from the hypotheses based on X (H_0^X) being the effect whereas the lower panels represent the results from the hypotheses based on Y (H_0^Y) being the effect. The employed measures of divergence are displayed in the legends.

Figure 8: **Functional time-series and the neighbourhood-based divergence.** The results of the functional time-series simulations as outlined in **Functional time-series**. The bivariate scatter plots demonstrate the functional relationship between the two variables at each lag and for each noise level. The plotted p-values are associated with the Granger-Sargent test and the neighbourhood-based normalized conditional mutual information.

Figure 9: **Multivariate time-series results using the Jensen-Shannon Divergence.** The aggregate results of the simulated multivariate time-series as outline in **Multivariate time-series**. The measure of interest here is the Jensen-Shannon divergence. The results are represented in a lag-wise manner as each lag stretching further back into the 'past' reveals a stronger causal relationship on aggregate. The color-coding of the arrows is based on the spearman correlation coefficient of the most recently included lag with the response variable/the effect.

Figure 10: **Multivariate time-series results using the normalized conditional mutual information.** *The aggregate results of the simulated multivariate time-series as outline in **Multivariate time-series**. The measure of interest here is the normalized conditional mutual information. The results are represented in a lag-wise manner as each lag stretching further back into the 'past' reveals a stronger causal relationship on aggregate. The color-coding of the arrows is based on the spearman correlation coefficient of the most recently included lag with the response variable/the effect.*