Methodology and Computing in Applied Probability (2022) 24:2587–2632
https://doi.org/10.1007/s11009-022-09928-3

Statistical Causality for Multivariate Nonlinear Time Series via Gaussian Process Models

Anna B. Zaremba[1] · **Gareth W. Peters**[2]

Received: 21 August 2020 / Revised: 22 September 2021 / Accepted: 4 January 2022 /
Published online: 30 March 2022

Abstract
The ability to test for statistical causality in linear and nonlinear contexts, in stationary or non-stationary settings, and to identify whether statistical causality influences trend of volatility forms a particularly important class of problems to explore in multi-modal and multivariate processes. In this paper, we develop novel testing frameworks for statistical causality in general classes of multivariate nonlinear time series models. Our framework accommodates flexible features where causality may be present in either: trend, volatility or both structural components of the general multivariate Markov processes under study. In addition, we accommodate the added possibilities of flexible structural features such as long memory and persistence in the multivariate processes when applying our semi-parametric approach to causality detection. We design a calibration procedure and formal testing procedure to detect these relationships through classes of Gaussian process models. We provide a generic framework which can be applied to a wide range of problems, including partially observed generalised diffusions or general multivariate linear or nonlinear time series models. We demonstrate several illustrative examples of features that are easily testable under our framework to study the properties of the inference procedure developed including the power of the test, sensitivity and robustness. We then illustrate our method on an interesting real data example from commodity modelling.

Keywords Statistical causality · Granger causality · Generalised likelihood ratio test · Nested models · ARD kernel

Mathematics Subject Classification 60G15 · 62F03 · 62M10

✉ Anna B. Zaremba
anna.b.zaremba@gmail.com

Gareth W. Peters
garethpeters@ucsb.edu

[1] Department of Computer Science, University College London, London WC1E 6EA, UK

[2] Janet & Ian Duncan Endowed Chair of Actuarial Science, Chair Professor of Statistics for Risk and Insurance, Department of Statistics and Applied Probability, University of California Santa Barbara, Santa Barbara, USA

1 Introduction

There are multiple notions of causality present in the statistics, econometrics and machine learning literature. We will consider one of these which is widely known as the class of causal concepts termed "statistical causality". We therefore, do not enter into any additional debate about merits of or frameworks for other notions of causality that may be common in areas of structured learning. Quoting Wiener (1956) "*For two simultaneously measured signals, if we can predict the first signal better by using the past information from the second one than by using the information without it, then we call the second signal causal to the first one.*" The general concept of statistical causality is based on comparing two predictive models, and this will be the case regardless of the type of predictive models used.

We seek to define a class of causality tests which is very general, and in principle agnostic to the class of underlying process that generates the time series being studied. We will achieve this via a class of semi-parametric models that we will utilise to model structural hypotheses regarding how causality may have manifested in the observed vector valued processes. To characterise testing of such relationships for the specific class of models we will develop, we will be able to explicitly evaluate a test statistics for a hypothesis test in which the asymptotic distribution is known in closed form, and under conditions discussed it can be shown to be the uniformly most powerful test. To achieve this we will introduce the class of representations we develop to characterise the observed vector valued time series according to Gaussian process models. These models are flexible in that they will efficiently allow us to test very general linear and nonlinear causality structures in the trend or volatility dynamics of the observed time series.

Throughout, we will consider without loss of generality, three multivariate time series denoted generically by $\mathbf{X}_t \in \mathbb{R}^p, \mathbf{Y}_t \in \mathbb{R}^{p'}$ and $\mathbf{Z}_t \in \mathbb{R}^{\bar{p}}$, which will be treated as column vectors. Our goal is to develop a framework in order to be able to assess conjectures regarding temporal relationships between these two multivariate processes $\mathbf{X}_t, \mathbf{Y}_t$, in the presence of side information $\mathbf{Z}_t$, such that we can apply formal inference procedures to determine the strength of evidence for or against such conjectures. We aim to achieve this in as general a manner as possible in order to accommodate differing forms of these relationships such as linear and nonlinear as well as stationary or non-stationary relationships. The approach we adopt will not require specific assumptions on the mechanism or model that generated these two series, which is important to understand. We form a distinction between the models used to postulate and test for the presence or absence of relationships between processes in a statistical causal manner and the knowledge of the true model or data generating processes. We will propose a framework which is applicable to testing very flexible and general relationships of causality whilst still potentially being misspecified with respect to the true data generating mechanism. As such, we note that we do not seek to perfectly represent the true underlying data generating processes, we simply seek to determine plausible relationships and existence of different causality relationships. This can be seen as a more general result than trying to model exactly the true underlying processes of the time series, as our approach can be used for rapid screening of multiple hypotheses about causality prior to a more detailed model development procedure.

To facilitate such generality, we set up the testing framework based on a model of the time series causal relationships captured by Gaussian process. Note, here we are not assuming the data is necessarily truly generated by a Gaussian process, but rather that the relationships of causality may be adequately reflected by such processes. This therefore only assumes a smooth variation of the causal relationships between the partially observed time

series represented by data $\{\mathbf{X}_t\}, \{\mathbf{Y}_t\}, \{\mathbf{Z}_t\}$. One particularly advantageous feature of working with Gaussian process representations of the causal relationships that are conjectured to be expressed by the time series is that we are able to derive and efficiently calculate relevant test statistics to perform inference of relevance to detection of causality structures in both linear and nonlinear classes.

1.1 Perspectives on Causal Analysis

The concept of statistical causality central to this paper is only one of numerous concepts of causality that have been proposed. For centuries, causality was studied by philosophers, until the advancements in science generated a need to express this concept in mathematical terms. As a consequence, there turns out to be a wide array of possible mathematical ways to express the concepts inherent in causality. In this section we would like to pay special attention to the General Theory of Causation by Pearl, and explain how Granger's statistical causality and Pearl's theory of causation cater to different needs.

Granger made certain assumptions, that he has called axioms Granger (1980), which are still central to statistical causality:

A1 **Time ordering:** states that the cause happens prior to the effect;
A2 **No redundant information:** states that the cause contains unique information about the effect – it is not related via a deterministic function.
A3 **Consistency:** says that the existence and direction of the causal relationship remain constant in time.

We note that Granger has also pointed out the contentious nature of this third axiom "[...] *generally accepted, even though it is not necessarily true*", which he saw as central to the applicability of the concept of causality.

However, it is important to understand that causal theory as developed by Pearl (2000, 2010) does not recognise these axioms. Instead Pearl postulated that causal analysis should allow inferring probabilities under static conditions, as well as how they change under dynamic conditions, by answering three types of questions:

Q1 Policy evaluation: What is the effect of potential intervention?
Q2 Probabilities of counterfactuals: Can an event be identified as responsible for another event?
Q3 Mediation: Can causal effect be assessed as direct or indirect?

Furthermore, Pearl clearly distinguished between associational and causal concepts: "*An associational concept is any relationship that can be defined in terms of a joint distribution of observed variables, and a causal concept is any relationship that cannot be defined from the distribution alone*" Pearl (2010). Following that criterion, causality in the sense of Granger – which here and in literature is referred to as either "statistical causality", or "Granger causality" – is an associational concept that is conditional and probabilistic in nature. According to Pearl, given adequately large sample and precise measurements, one can in principle test associational assumptions, but not causal assumptions, which require experimental control. The last one is of crucial importance to understanding the differences in the applicability of those two concepts. Statistical causality has been developed for, and

is often used in studying time series, and in that context no experimental control is available in an observational context rather than a designed experimental context.

We conclude this part by referring the reader to the literature that offers tools for reconciling statistical causality with Pearl's General Theory of Causation. Eichler (2001) proposed using Granger causality graphs, and Eichler and Didelez (2007) introduced interventions to the Granger causal framework in a way that is similar to Pearl's approach, and thus offer tools for reconciling statistical causality with Pearl's General Theory of Causation. White et al. (2011) demonstrated how Pearl's Causal Model and Granger causality are linked when expressed in terms of extension of Pearl's Causal Model with settable systems.

1.2 Contributions

In this section we briefly outline the novelty of the proposed statistical causality framework developed and contrast the contributions to related references Amblard et al. (2012a, b). In Amblard et al. (2012a) they propose the use of Gaussian Process (GP) models for univariate time series studies and use the data evidence obtained from the models to design a test for causality. However, these works do not explore the complete flexibility of these classes of models, where one can generalise readily to multivarite time series settings, also add side information, and they do not explore the range of causal structures that includes linear or nonlinear structures in the trend or covariance, as well as the estimation optimisation procedure, and model calibration aspects that we generalise in this manuscript. Furthermore, no sensitivity studies or analysis of the effect of model misspecifications is undertaken in the aforementioned works. Despite formulating test statistic as a difference in marginal loglikelihoods, Amblard et al. (2012a) do not exploit the properties of likelihood ratio type tests (LRT or GLRT). Therefore, in this manuscript we build upon these interesting papers and we genearlise the class of models significantly as they have not formulated the required nested structures that are achieved in this manuscript to facilitate more direct hypothesis testing frameworks through the LRT and GLRT, where we may take advantage of well known properties of the resulting test statistics asymptotic distribution under the null.

We would also like to point out that in our research we found the "kernelised Geweke's measure" of Amblard et al. (2012a) to be a nonlinear generalisation of Granger causality with many good properties, but with several shortcomings that we were able to successfully address by the use of GPs (notably: the difficulty of hyperparameter estimation, model calibration, lack of interpretability of model parameters). However the authors of Amblard et al. (2012a) have progressed from what we see as more general model (GP) to a less general model (kernelised ridge regression), which they saw as more practical for modelling instantaneous causality Amblard et al. (2012b). In our manuscript we do not address instantaneous causality (instantaneous coupling), but our framework can incorporate it, just like it can model causality between multivariate time series.

The multivariate causal testing framework developed in this manuscript allows one to incorporate aspects of causality, linear and nonlinear, in the mean and the covariance. In line with the very general definition of non-causality, models of statistical causality typically test for the equivalence of two conditional distributions. One can then differentiate approaches based on what further assumptions are made on the models. For instance, linear regression methods focus on recognising dependence in trend, under strict model assumptions, while nonlinear generalisations relax these model assumptions. These models do not, however, allow for causality in covariance, or any other

nonlinear structures. The framework developed in this manuscript can accomodate these valuable extensions to allow direct straightforward testing of causality in the covariance in linear and nonlinear settings.

Secondly, analysing causal structure with Gaussian processes hasn't been done in the likelihood ratio framework, we suppose due to the complication in formulating a nested testing model structure. In this manuscript we propose a way to construct model nesting that allows for application of the likelihood ratio test (LRT) and Generalised Likelihood Ratio Test (GLRT). This model nesting is constructed to be applicable for assessing causality in the mean, or covariance, or both, and is achieved through Automatic Relevance Determination (ARD) construction of the kernel. The development of nested models is important, as the standard asymptotic distribution of the LRT test statistic under the null being χ^2 does not hold for non-nested hypotheses. Thus, we emphasise that the novelty does not lie in the development of the asymptotic behaviour of the test, which is standard, but in constructing a framework that allows to apply that test in this general and flexible statistical causality framework we propose. Furthermore, with our GP model formulations the test statistic can be written in a closed form, can be computed point-wise, and is efficient to compute.

There are numerous advantages of using GPs, beginning with: ease of optimisation and interpretability of hyperparameters, flexibility, richness of covariance functions, allowing for various model structures. Using a likelihood ratio type test with a GP is a very natural choice, as estimating GP model parameters is often done on the basis of maximising likelihood, and therefore this estimation can be incorporated into the compound version of the likelihood ratio test (Generalised Likelihood Ratio Test, GLRT). From Gaussian variables, GPs inherited the property of being fully specified by the mean and the covariance, and so testing for model equivalence inherently means testing for equivalence of the mean and covariance functions. But many popular kernels do not have the ARD property, and using them for a likelihood ratio test settings gives no easy way to account for causal structures in covariance. Consequently, it is using GLRT with an ARD-GP that gives a uniformly most powerful test with an unparalleled flexibility: known asymptotic distribution under the null, explicit evaluation and in a closed form, and usefulness also for misspecified models.

Thirdly, we demonstrate the ability to detect and identify causal structures in the mean and covariance, even in the presence of different types of model misspecifications. We undertake careful study of sensitivity and robustness of these testing frameworks to various features that one would encounter, like: sample size, parameter misspecification and structural misspecification. It is important as these studies demonstrate that one can reliably apply these tests in a general framework, even if the model is misspecified in those ways, and still have confidence that the inference procedure can detect these types of causality in mean and covariance incorporated in this framework reliably.

2 Concepts in Statistical Causality

It is important to understand the context of our proposed framework in light of the specific formulations that have been proposed before when studying the concept of statistical causality. We, therefore, briefly outline previous approaches to consider this testing framework for statistical causality and importantly the required assumptions.

2.1 Granger Causality

In this section, we present the original formulation of the hypothesis tests and test statistics for Granger causality, as well as a few of their later extended formulations that form the basis for considering concepts of statistical causality.

The first testable form of statistical causality proposed by Granger (1963) was developed in the context of linear forms of vector autoregressive models. For time series $\{\mathbf{X}_t\}, \{\mathbf{Y}_t\}$ and $\{\mathbf{Z}_t\}$ lets assume we consider two alternative model formulations for $\{\mathbf{Y}_t\}$:

$$\text{Model A:} \qquad \mathbf{Y}_t = \sum_{j=1}^{l} A_{22,j}\mathbf{Y}_{t-j} + \sum_{j=1}^{l} A_{23,j}\mathbf{Z}_{t-j} + \epsilon_{Y,t}, \tag{1}$$

$$\text{Model B:} \qquad \mathbf{Y}_t = \sum_{j=1}^{l} A_{21,j}\mathbf{X}_{t-j} + \sum_{j=1}^{l} A'_{22,j}\mathbf{Y}_{t-j} + \sum_{j=1}^{l} A'_{23,j}\mathbf{Z}_{t-j} + \epsilon'_{Y,t}, \tag{2}$$

with $l \in \mathbb{N}$ being the maximum number of lagged observations and $\epsilon_{Y,t}, \epsilon'_{Y,t}$ denoting noise (later E^X_t, E^Y_t will denote residuals). In this setting, we introduce Granger's definition:

Definition 1 Granger (1963) Causality of the process Y by the process X is defined when: $Var(E'_{Y,t}) < Var(E_{Y,t})$, and denoted as: $X \to Y$. There is no causality, if $Var(E'_{Y,t}) = Var(E_{Y,t})$, and this is denoted as: $X \nrightarrow Y$.

Typically the hypotheses will be given by one of the following sets of null hypothesis of Granger non-causality and alternative hypothesis of lack of Granger non-causality[1]. The version 1 of the hypothesis of non-causality is consistent with the Definition 1:

$$\begin{aligned} \text{version 1} \qquad H_0 &: \quad Var(E'_{Y,t}) = Var(E_{Y,t}), \\ H_1 &: \quad Var(E'_{Y,t}) < Var(E_{Y,t}). \end{aligned} \tag{3}$$

In the specific case of the linear regression models from the Eqs. 1 and 2, we can also use the version 2 of the hypothesis of non-causality (which implies version 1):

$$\begin{aligned} \text{version 2} \qquad H_0 &: \quad \forall j \in \{1, .., l\} \;\; A_{21,j} = 0, \\ H_1 &: \quad \exists j \in \{1, .., l\} \;\; A_{21,j} \neq 0. \end{aligned} \tag{4}$$

Granger proposed to test the null hypothesis from Eq. 3 using a test statistic called **strength of causality** and denoted $L^{SC}_{X\to Y}$, defined as the ratio of the two variances of prediction errors:

$$L^{SC}_{X\to Y} = 1 - \frac{Var(E'_{Y,t})}{Var(E_{Y,t})}, \quad \text{where} \;\; 0 \leq L^{SC}_{X\to Y} \leq 1. \tag{5}$$

The strength of causality underlines the relationship between Granger causality and model specification tests for linear regression.

Since this instrumental work there have been numerous developments and extensions proposed. For instance Geweke (1982) proposed **measure of linear feedback**, with the same model assumptions and equivalent to strength of causality (Eq. 5), and defined as:

[1] We never test for existence of causality, but only accept or reject the hypothesis of lack of causality.

$$L_{X\to Y}^{MLF} = \ln\left(\frac{\mid Var(E_{Y,t}) \mid}{\mid Var(E'_{Y,t}) \mid}\right), \tag{6}$$

which will be χ^2_p distributed under the null hypothesis of lack of causality.

Kernelised version of the Geweke's measure of linear feedback has been proposed in Amblard et al. (2012b) with the aim to make it a nonlinear method. This kernelised version of the measure of linear feedback has the same form as in the Eq. 6, but arises from a different model: kernel ridge regression, with the best predictor in the reproducing kernel Hilbert space (RKHS) generated by the associated kernel.

In the kernel ridge regression the solution is no longer represented in terms of optimised coefficients $A_{\cdot\cdot,j}$ from Eqs. 1 and 2, the so called **primal solution** which we will denote as $\boldsymbol{\alpha}$. Instead, the **dual solution** $\boldsymbol{\beta}^{krr}$ are such coefficients, that allow the solution to be represented in terms of inner product of the covariates (independent variables). Below, we introduce notation that will allow convenient matrix operations, and that will be used throughout also in the later sections:

$$\begin{aligned}
&\mathbf{Y}_t \in \mathbb{R}^{p'}, && p' \times 1 \text{ column vector}\\
&\mathbf{Y}_{1:T} := \left[\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_T\right]^{\mathrm{T}}, && T \times p'\\
&\mathbf{Y}_t^{-l} := \left[\mathbf{Y}_{t-l+1}^{\mathrm{T}}, \mathbf{Y}_{t-l+2}^{\mathrm{T}}, ..., \mathbf{Y}_t^{\mathrm{T}}\right], && 1 \times (lp')\\
&\mathbf{Y}^{-l} := \mathbf{Y}_{1:T}^{-l} = \left[\mathbf{Y}_{1-l+1:T-l+1}, \mathbf{Y}_{1-l+2:T-l+2}, ..., \mathbf{Y}_{1:T}\right], && T \times (lp')\\
&\mathbb{Q}_t := \left[\mathbf{X}_t^{\mathrm{T}}, \mathbf{Y}_t^{\mathrm{T}}, \mathbf{Z}_t^{\mathrm{T}}\right] \text{ for model B}, && 1 \times (p + p' + \bar{p})\\
&\mathbb{Q} := \left[\mathbf{X}^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m}\right] \text{ for model B}, && T \times (kp + lp' + m\bar{p})
\end{aligned}$$

For Model A, we have $\mathbb{Q}_t := \left[\mathbf{Y}_t, \mathbf{Z}_t\right], \mathbb{Q} := \left[\mathbf{Y}^{-l}, \mathbf{Z}^{-m}\right]$, and if these two need to be distinguished, we will add subscript referring to the model: $\mathbb{Q}_{B,t}, \mathbb{Q}_{A,t}$. Here the matrix $\mathbb{Q}$ represents all available covariate data, and the matrix $\mathbb{Q}\mathbb{Q}^{\mathrm{T}} : T \times T$ is a Gramm matrix of the covariates data – it is in such a form that admits application of the kernel trick by "substituting" it with a Gramm matrix which we will denote $\mathbf{K}_{\mathbb{Q}}$. The Gramm matrix $\mathbf{K}_{\mathbb{Q}}$, also called kernel matrix or covariance matrix, can be defined element-wise as Mercer kernel function evaluations: $\left\{\mathbf{K}_{\mathbb{Q}}\right\}_{i,j} = k(\mathbf{Q}_{i-l:i-1}, \mathbf{Q}_{j-l:j-1})$. The Mercer kernel function $k(\cdot,\cdot) \in M(\mathcal{X})$ is a real, symmetric and semi-positive definite kernel function, defined on the domain $\mathcal{X} \times \mathcal{X}$. Then, the optimal weights, fitted values and mean square of prediction error will for kernel ridge regression be as follows:

$$\begin{aligned}
&\text{optimal weights:} && \boldsymbol{\beta}^{krr} = \left(\mathbf{K}_{\mathbb{Q}} + \lambda \mathbf{1}_{T-1}\right)^{-1} \mathbf{Y}_{1:T}\\
&\text{fitted values:} && \hat{\mathbf{Y}}_{1:T} = \mathbf{K}_{\mathbb{Q}} \boldsymbol{\beta}^{krr}\\
&\text{MSE:} && Var\left(\hat{\mathbf{Y}}_{1:T} - \mathbf{Y}_{1:T}\right) = \frac{1}{T-1}\left(\mathbf{K}_{\mathbb{Q}} \boldsymbol{\beta}^{krr} - \mathbf{Y}_{1:T}\right)^{\mathrm{T}} \left(\mathbf{K}_{\mathbb{Q}} \boldsymbol{\beta}^{krr} - \mathbf{Y}_{1:T}\right).
\end{aligned}$$

When kernel ridge regression is applied to model A, or model B, all of the steps above are applied, but with different definition of Q_t, and therefore different values of the covariance matrix $\mathbf{K}_{\mathbb{Q}}$. Denoting the fitted values as $\hat{\mathbf{Y}}^A_{1:T}$ and $\hat{\mathbf{Y}}^B_{1:T}$, we obtain the mean square errors of kernel ridge regression prediction of the two models: $Var\left(\hat{\mathbf{Y}}^A_{1:T} - \mathbf{Y}_{1:T}\right)$ and $Var\left(\hat{\mathbf{Y}}^B_{1:T} - \mathbf{Y}_{1:T}\right)$, which are used in the test statistic in a similar manner to the strength of causality from Eq. 5, and to the test statistic from Eq. 6. Thus

the test statistic based on the kernelised ridge regression, that Amblard et al. (2012b) proposed is formulated as follows:

$$L^{krr}_{X\to Y} = \ln \frac{Var\left(\hat{\mathbf{Y}}^{A}_{1:T} - \mathbf{Y}_{1:T}\right)}{Var\left(\hat{\mathbf{Y}}^{B}_{1:T} - \mathbf{Y}_{1:T}\right)}. \tag{7}$$

The hypotheses are:

$$\begin{aligned} H_0 &: L^{krr}_{X\to Y} = 0, \quad \text{no causality from } \{X\} \text{ to } \{Y\} \\ H_1 &: L^{krr}_{X\to Y} > 0, \quad \text{causality from } \{X\} \text{ to } \{Y\} \end{aligned}$$

See Amblard et al. (2012b); Zaremba and Aste (2014). We also refer to Lungarella et al. (2007) for other generalisations of Granger causality.

2.2 Transfer entropy

A third set of hypothesis has been subsequently introduced, which also relies on concepts of conditional independence, see Granger (1980); Eichler (2001); Amblard et al. (2012b):

$$\begin{aligned} \text{version 3} \qquad H_0 &: \quad p(\mathbf{Y}_t \mid \mathbf{X}^{-k}_{t-1}, \mathbf{Y}^{-l}_{t-1}, \mathbf{Z}^{-m}_{t-1}) = p(\mathbf{Y}_t \mid \mathbf{Y}^{-l}_{t-1}, \mathbf{Z}^{-m}_{t-1}), \quad \forall t \in \mathbb{Z} \\ H_1 &: \quad p(\mathbf{Y}_t \mid \mathbf{X}^{-k}_{t-1}, \mathbf{Y}^{-l}_{t-1}, \mathbf{Z}^{-m}_{t-1}) \neq p(\mathbf{Y}_t \mid \mathbf{Y}^{-l}_{t-1}, \mathbf{Z}^{-m}_{t-1}), \quad \forall t \in \mathbb{Z}. \end{aligned} \tag{8}$$

The hypotheses in Eq. 8 were a starting point for a wide range of other tests, many of which would no longer assume the linear form of the models in the Eqs. 1 and 2, see Schreiber (2000); Lungarella et al. (2007); Chen (2006). One of the more important papers here is the one by Schreiber (2000) who introduced the information theoretic approach to modelling causality by proposing transfer entropy, which is now one of the most popular nonlinear statistical causality measures. Transfer entropy is defined as a difference of two conditional entropies:

$$L^{TE}_{X\to Y} = H(\mathbf{Y}_t \mid \mathbf{Y}^{-l}_{t-1}) - H(\mathbf{Y}_t \mid \mathbf{X}^{-k}_{t-1}, \mathbf{Y}^{-l}_{t-1}),$$

where

$$H(\mathbf{X}) = -\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x}$$

is the differential (continuous) entropy and

$$H(\mathbf{X} \mid \mathbf{Y}) = \int \int p(\mathbf{x}, \mathbf{y}) \ln \frac{p(\mathbf{x})}{p(\mathbf{x}, \mathbf{y})} d\mathbf{x} d\mathbf{y}$$

is the conditional entropy.

The asymptotic properties of transfer entropy are analysed less often. However Barnett and Bossomaier (2012) prove that for ergodic processes, the transfer entropy is a log-likelihood ratio, asymptotically distributed according to the χ^2 distribution under the null hypothesis of lack of causality, and having asymptotic non-central χ^2 distribution for the alternative hypothesis. The rate of convergence of the test statistics asymptotic distribution can however be problematic in practice, requiring very large sample sizes for valid

application of the test according to the asymptotic distribution (please refer to the experimental results, Sect. 5.6).

In general, the null hypothesis from Eq. 8 is not equivalent to neither that from Eq. 3 nor from Eq. 4. For the linear model from Eqs. 1 and 2, and with the assumptions of $\epsilon_{Y,t}, \epsilon'_{Y,t}$ having the same distributions, the null hypothesis from Eq. 4 implies the null hypothesis from Eq. 8. Furthermore, under the normality assumptions, the test statistic $L^{TE}_{X\to Y}$ is equivalent to both $L^{SC}_{X\to Y}$ and $L^{MLF}_{X\to Y}$, see Barnett and Bossomaier (2012).

3 Semi-Parametric Nonlinear Causal Process Representations

We begin by defining Gaussian Processes, as this will serve as our base class of stochastic processes that we adopt to characterise different examples of causality model structures. The vector valued time series $\{\mathbf{Y}_t\}$ is described by a Gaussian Process model, which is denoted as $\mathcal{GP}$ and defined as follows:

Definition 2 (Gaussian Process (GP)) Denote by $f(\mathbf{x}) : \mathcal{X} \mapsto \mathbb{R}$ a stochastic process parametrised by $\{\mathbf{x}\} \in \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{R}^p$. Then, the random function $f(\mathbf{x})$ is a Gaussian process if all its finite dimensional distributions are Gaussian, where for any $N \in \mathbb{N}$, the random vector $\big(f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_N)\big)$ is jointly normally distributed, see Rasmussen and Williams (2006).

We can therefore interpret a GP as formally defined by the following class of random functions:

$$
\begin{aligned}
f := \{ & f(\cdot) : \mathcal{X} \mapsto \mathbb{R}, \text{ s.t. } f(\cdot) \sim \mathcal{GP}\big(\mu(\cdot;\boldsymbol{\theta}_\mu), k(\cdot,\cdot;\boldsymbol{\theta}_k)\big), \text{ with} \\
& \mu(\cdot;\boldsymbol{\theta}_\mu) := \mathbb{E}\big[f(\cdot)\big] : \mathcal{X} \mapsto \mathbb{R}, \\
& k(\cdot,\cdot;\boldsymbol{\theta}_k) := \mathbb{E}\big[\big(f(\cdot) - \mu(\cdot;\boldsymbol{\theta}_\mu)\big)\big(f(\cdot) - \mu(\cdot;\boldsymbol{\theta}_\mu)\big)\big] : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}^+ \}.
\end{aligned}
$$

At each point the mean of the function is $\mu(\cdot;\boldsymbol{\theta}_\mu)$, parametrised by $\boldsymbol{\theta}_\mu$, and the dependence between any two points is given by the covariance function, also called Mercer kernel: $k(\cdot,\cdot;\boldsymbol{\theta}_k) : M(\mathcal{X})$, parametrised by $\boldsymbol{\theta}_k$, see detailed discussion in Rasmussen and Williams (2006). We will later use notation $\boldsymbol{\theta} = \boldsymbol{\theta}_\mu \cup \boldsymbol{\theta}_k$, and will refer to $\boldsymbol{\theta}$ as hyperparameters of the GP random function f.

We then model the time series $\{\mathbf{Y}_t\}$ causal relationships as realisations[2] from a GP $f(\cdot)$ with additive Gaussian noise ϵ_t.

$$
\mathbf{Y}_t = f(\mathbf{X}^{-k}_{t-1}, \mathbf{Y}^{-l}_{t-1}, \mathbf{Z}^{-m}_{t-1}) + \epsilon_t, \quad f(\cdot) \sim \mathcal{GP}\big(\mu_t, k_{t,s};\boldsymbol{\theta}\big), \quad \epsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2),
$$

with the following generic definition of the mean function $\mu_t : \mathbb{R}^{kp+lp'+m\bar{p}} \to \mathbb{R}$ and the covariance function $k_{t,s} : \mathbb{R}^{kp+lp'+m\bar{p}} \times \mathbb{R}^{kp+lp'+m\bar{p}} \to \mathbb{R}$:

[2] For multivariate model one can use multiple output GP: please refer to Cressie (1993) for literature about "cokriging", to Boyle and Frean (2005); Alvarez and Lawrence (2011) for multiple output GP processes modelled as convolutions of the same underlying white noise process.

$$\mu_t := \mu_t\left(\left[\mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}\right]\right),$$
$$k_{t,s} := k_{t,s}\left(\left[\mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}\right], \left[\mathbf{X}_{s-1}^{-k}, \mathbf{Y}_{s-1}^{-l}, \mathbf{Z}_{s-1}^{-m}\right]\right),$$

It will be useful to make the following notational definitions for the mean vector, and correlation matrix, respectively:

$$\boldsymbol{\mu} := \left[\mu_1, \ldots, \mu_T\right]^{\mathrm{T}}, \quad \mathbf{K} := \begin{bmatrix} k_{1,1} & \cdots & k_{1,T} \\ \vdots & \ddots & \vdots \\ k_{T,1} & \cdots & k_{T,T} \end{bmatrix}, \mathbf{K} \in SPD_T,$$

and SPD_T is the manifold of symmetric positive definite matrices of size $T \times T$.

3.1 Covariance Functions and Automatic Relevance Determination for Causality

As is standard in GP modelling, we will represent the covariance functions with functions that are known as kernels, and we will focus on the class of Mercer kernels $M(\mathcal{X})$.

Definition 3 (Semi-positive definite kernel) A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a semi-positive definite kernel kernel (positive definite) if and only if it is symmetric, that is, $\forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}, k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}', \mathbf{x})$ and semi-positive definite, that is

$$\forall \mathbf{x}_1, ..., \mathbf{x}_N \in \mathcal{X} \quad \forall c_1, ..., c_N \in \mathbb{R} \quad \sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j k(\mathbf{x}_i, \mathbf{x}_j) \geqslant 0.$$

There are several important properties of kernels, see Scholkopf and Smola (2001). A centered GP is uniquely determined by its covariance function (semi-positive definite kernel). Conversely, any semi-positive definite kernel determines a covariance function and a unique centered GP, see Hein and Bousquet (2004). Moreover, there exists a bijection between the set of all real-valued semi-positive kernels on some space $\mathcal{X}$ and the set of all centered GPs defined on $\mathcal{X}$. Kernels can also be seen as inner products, see Schoelkopf et al. (2004).

An important concept that will be broadly used in the context of kernel classes is the concept of **Automatic Relevance Determination** (ARD). It has been initially introduced by MacKay (1994), as a Bayesian model where input relevance can be introduced and controlled with parameters; see also Neal (1996). This has later become popular in a wider context of feature selection and sparse learning in Bayesian models, see Qi et al. (2004). We use the same concept, but for a purpose of ensuring we have nested models for inference hypothesis design (see Sect. 4.1.3), and it will be crucial when applying the Generalised Likelihood Ratio Test.

In the ARD model, each input variable has an associated hyperparameter whose value can scale the effect of that input. In the Bayesian approach, this is achieved by setting a separate Gaussian prior for each of the inputs. In our (frequentist) case we treat each dimension as a separate input and define our mean and covariance functions in such a manner that the effect of each of the univariate inputs can be separately changed through zeroing of the hyperparameter associated with the given marginal input component. In particular, by setting specific values of the hyperparameters we can practically eliminate some of the univariate variables from the mean/covariance. This construction has several important advantages: it allows for marginal causality testing as well as developing a class of nested model structures, critical to determining the statistical

Table 1 Summary of several popular kernel functions. We are using the following notation: $p_1 \leq p$ is the dimension of vectors $\mathbf{x}_u, \mathbf{x}_v$, and $\mathbf{x}_{u,[1:p_1]}^T = [x_{u,1}, ..., x_{u,p_1}]$, A is a constant positive definite matrix, a, c are a constants, l is a lengthscale parameter, and $[l_1, ..., l_p]$ is a vector of lengthscale parameters, $d = ||\mathbf{x}_u - \mathbf{x}_v||$ represents a distance, e.g. an Euclidean distance, $\mathbf{D} = [||x_{u,1} - x_{v,1}||, ..., ||x_{u,p} - x_{v,p}||]$, and $k_1(\cdot), k_2(\cdot)$ are stationary kernels

covariance function	expression $k(\mathbf{x}_u, \mathbf{x}_v) =$	stationarity	hyparameter domains
constant (noise)	$diag(\sigma_1^2, ..., \sigma_p^2)\delta_{\mathbf{x}_u \mathbf{x}_v}$	+	$\sigma_i^2 \in \mathbb{R}^+$
linear	$\mathbf{x}_u^T \mathbf{x}_v$	-	
linear ARD	$\mathbf{x}_u^T A \mathbf{x}_v$	-	$A \in PSD_p$
polynomial	$\sigma_f^2 (a + \mathbf{x}_u^T \mathbf{x}_v)^c$	-	$\sigma_f^2 \in \mathbb{R}^+,\ a \in \mathbb{R},$ $c \in \mathbb{N} \cup 0$
squared exponential	$\sigma_f^2 \exp\left(-\frac{(\mathbf{x}_u - \mathbf{x}_v)^T(\mathbf{x}_u - \mathbf{x}_v)}{2l^2}\right)$	+	$\sigma_f^2 \in \mathbb{R}^+, l \in \mathbb{R}^+$
squared exponential ARD	$\sigma_f^2 \exp\left(-\frac{1}{2}\mathbf{D}\,\mathrm{diag}\left(\left[l_1^2, ..., l_p^2\right]\right)\mathbf{D}^T\right)$	+	$\sigma_f^2 \in \mathbb{R}^+, l_i \in \mathbb{R}^+ \cup 0$
Matern	$\frac{\sigma_f^2}{\Gamma(\nu)2^{\nu-1}}\left(\sqrt{2\nu}\frac{d}{l}\right)^{\nu} K_\nu\left(\sqrt{2\nu}\frac{d}{l}\right)$	+	$\sigma_f^2 \in \mathbb{R}^+,\ \nu \in \mathbb{R},$ $l \in \mathbb{R}^+$
Matern ARD	$\frac{\sigma_f^2}{\Gamma(\nu)2^{\nu-1}}\left(\sqrt{2\nu}\mathbf{D}[l_1, ..., l_p]^T\right)^{\nu} K_\nu\left(\sqrt{2\nu}\mathbf{D}[l_1, ..., l_p]^T\right)$	+	$\sigma_f^2 \in \mathbb{R}^+,\ \nu \in \mathbb{R}^+,$ $l_i \in \mathbb{R}^+ \cup 0$
periodic	$\sigma_f^2 \exp\left(-2\sum_{i=1:p} \frac{\sin(\pi(\mathbf{x}_{v,i} - \mathbf{x}_{u,i})/a)}{l_i^2}\right)$	+	$\sigma_f^2 \in \mathbb{R}^+, l_i \in \mathbb{R}+$
separable nonstationary	$k_1\left(\mathbf{x}_{u,[1:p_1]}, \mathbf{x}_{v,[1:p_1]}\right) k_2\left(\mathbf{x}_{u,[p_1+1:p]}, \mathbf{x}_{v,[p_1+1:p]}\right)$	-	$p_1 \leq p$

significance of causality relationships under consideration. In the table below (Table 1) are two examples of popular kernels and their ARD versions. Rasmussen and Williams in their MATLAB toolbox provide an ARD version of the squared exponential kernel with $diag\left(\left[l_1^{-2}, ..., l_n^{-2}\right]\right)$, our version from the Table 1 allows to choose $l_i = 0$ which removes the effect of the i-th dimension of input on the kernel. As a result, the covariance for lower dimensional space can be expressed as a covariance with a higher dimensional space $k_{t,s}^{SE}([\mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}], [\mathbf{Y}_{s-1}, \mathbf{Z}_{s-1}]) = k_{t,s}^{SE}([\mathbf{X}_{t-1}, \mathbf{Y}_{t-1}, \mathbf{Z}_{t-1}], [\mathbf{X}_{s-1}, \mathbf{Y}_{s-1}, \mathbf{Z}_{s-1}]; l_1 = 0)$.

Given a set of input points $\{\mathbf{x}_i | i = 1, ..., n\}$ we can compute the Gram (covariance) matrix $\mathbf{K}$ whose entries are $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$.

4 Characterising Causality Hypotheses With Gaussian Process Models

When performing inferential tests for statistical causality one will typically compare two alternative model hypotheses. We have already seen in the Sect. 1, that such hypotheses can be formulated in multiple ways, see Eqs. 3, 4 and 8. In defining the non-causality tests, we start from the more general forms of the hypotheses outlined in Eq. 8.

The two causal model structures are generically represented as multi-dimensional Gaussian process time series models observed in additive Gaussian noise and denoted by Model A and Model B in the Eqs. 9 and 10 respectively:

$$\begin{aligned} \text{Model A: } \mathbf{Y}_t = f_A(\mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}) + \epsilon_t^A, \quad & f_A(\cdot) \sim \mathcal{GP}\big(\mu_{A,t}, k_{A,t,s}; \boldsymbol{\theta}_A, \mathcal{M}_A\big) \\ & \epsilon_t^A \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_A^2 \mathbb{I}_{p' \times p'}) \end{aligned} \tag{9}$$

$$\begin{aligned} \text{Model B: } \mathbf{Y}_t = f_B(\mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}) + \epsilon_t^B, \quad & f_B(\cdot) \sim \mathcal{GP}\big(\mu_{B,t}, k_{B,t,s}; \boldsymbol{\theta}_B, \mathcal{M}_B\big) \\ & \epsilon_t^B \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_B^2 \mathbb{I}_{p' \times p'}) \end{aligned} \tag{10}$$

with the following forms of mean functions $\mu_A : \mathbb{R}^{lp'+m\bar{p}} \to \mathbb{R}$, $\mu_B : \mathbb{R}^{kp+lp'+m\bar{p}} \to \mathbb{R}$ and covariance functions $k_A : \mathbb{R}^{lp'+m\bar{p}} \times \mathbb{R}^{lp'+m\bar{p}} \to \mathbb{R}$, $k_B : \mathbb{R}^{kp+lp'+m\bar{p}} \times \mathbb{R}^{kp+lp'+m\bar{p}} \to \mathbb{R}$:

$$\begin{aligned} \mu_{A,t} &:= \mu_A\big(\big[\mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}\big]\big), \\ k_{A,t,s} &:= k_A\big(\big[\mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}\big], \big[\mathbf{Y}_{s-1}^{-l}, \mathbf{Z}_{s-1}^{-m}\big]\big), \\ \mu_{B,t} &:= \mu_B\big(\big[\mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}\big]\big), \\ k_{B,t,s} &:= k_B\big(\big[\mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}\big], \big[\mathbf{X}_{s-1}^{-k}, \mathbf{Y}_{s-1}^{-l}, \mathbf{Z}_{s-1}^{-m}\big]\big). \end{aligned}$$

We assume the mean and covariance functions, μ_A, k_A and respectively μ_B, k_B, have similar functional forms and only differ in dimensionality and hyperparameters.

Having defined these two models we may now state the form of the hypotheses for testing for non-causality (lack of causality) in nonlinear times series. The test that allows comparing two models from the Eqs. 9 and 10 is fundamentally a test comparing two distributions – the conditional distribution of the time series $\{\mathbf{Y}_t\}$ conditioned on inputs from either of the two models. As it was already mentioned, we never actually confirm the statistical causality, but rather reject lack of causality (test for non-causality).

Under such a test, the null hypothesis is that there is no causal relationship from time series $\{\mathbf{X}_t\}$ to $\{\mathbf{Y}_t\}$, and including the past of $\{\mathbf{X}_t\}$ does not improve the prediction of $\{\mathbf{Y}_t\}$. Given the model formulations, this means equality of conditional distribution of $\mathbf{Y}$, conditioning on either set of explanatory variables (analogously to Eq. 8):

$$H_0 : \quad p(\mathbf{Y} \mid \mathbf{X}^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m}; \boldsymbol{\theta}_B, \mathcal{M}_B) = p(\mathbf{Y} \mid \mathbf{Y}^{-l}, \mathbf{Z}^{-m}; \boldsymbol{\theta}_A, \mathcal{M}_A) \tag{11}$$

$$H_1 : \quad p(\mathbf{Y} \mid \mathbf{X}^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m}; \boldsymbol{\theta}_B, \mathcal{M}_B) \neq p(\mathbf{Y} \mid \mathbf{Y}^{-l}, \mathbf{Z}^{-m}; \boldsymbol{\theta}_A, \mathcal{M}_A). \tag{12}$$

The distributions above can be obtained in closed form only in the case of additive Gaussian noise, or in cases where there is no assumed additive noise in Model A or model B.

Since a GP is also specified by its sufficient mean and covariance functions, testing for equality of distributions will be equivalent to testing for equality of the mean functions and the covariance functions. Hence, the convenient feature of the causality testing framework developed from the GP framework we propose is that these general distributional statements about population quantities in the null and alternative hypotheses are equivalent to the following population statements on mean and covariance functions.

$$
\begin{aligned}
H_0 : \quad & \exists k(.,.) \in M\left(\mathbb{R}^{lp'+m\bar{p}} \times \mathbb{R}^{lp'+m\bar{p}}\right), \mu : \mathbb{R}^{lp'+m\bar{p}} \to \mathbb{R}, \quad \forall t,s \in \{l+1,...,T\} \\
& k_B\left(\left[\mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}\right], \left[\mathbf{X}_{s-1}^{-k}, \mathbf{Y}_{s-1}^{-l}, \mathbf{Z}_{s-1}^{-m}\right]\right) \equiv k\left(\left[\mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}\right], \left[\mathbf{Y}_{s-1}^{-l}, \mathbf{Z}_{s-1}^{-m}\right]\right) \\
& \mu_B\left(\left[\mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}\right]\right) \equiv \mu\left(\left[\mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}\right]\right) \\
H_1 : \neg & \exists k(.,.) \in M\left(\mathbb{R}^{lp'+m\bar{p}} \times \mathbb{R}^{lp'+m\bar{p}}\right), \mu \in: \mathbb{R}^{lp'+m\bar{p}} \to \mathbb{R} \forall t,s \in \{l+1,...,T\} \\
& k_B\left(\left[\mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}\right], \left[\mathbf{X}_{s-1}^{-k}, \mathbf{Y}_{s-1}^{-l}, \mathbf{Z}_{s-1}^{-m}\right]\right) \equiv k\left(\left[\mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}\right], \left[\mathbf{Y}_{s-1}^{-l}, \mathbf{Z}_{s-1}^{-m}\right]\right) \\
& \mu_B\left(\left[\mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}\right]\right) \equiv \mu\left(\left[\mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}\right]\right),
\end{aligned}
$$

where M represents the class of all Mercer kernels.

If the classes of mean and covariance functions are restricted so that the Model A is nested in the Model B (defined in the Subsect. 4.1.3), then the above hypotheses can be tested with the Generalised Likelihood Ratio Test.

4.1 Generalised Likelihood Ratio Test

The GLRT is a composite hypothesis test that can be used in the case of nested hypothesis if the parameters are unknown and need to be estimated. Below we describe the test, using notation from Garthwaite et al. (2002). The GLRT gives us asymptotic distribution of the test statistics, but it requires that the hypotheses are nested – which can be expressed in terms of restriction on mean and covariance formulations.

4.1.1 Theory for Generalised Likelihood Ratio Test

Let $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_N$ be a random sample of size N from a distribution with pdf $p(\mathbf{x};\boldsymbol{\theta})$, and suppose that we wish to test: $H_0 : \boldsymbol{\theta} \in \omega$ vs $H_1 : \boldsymbol{\theta} \in \Omega - \omega$. Then define a random variable:

$$
\Lambda = \left\{ \max_{\boldsymbol{\theta} \in \omega} L(\boldsymbol{\theta};\mathbf{X}) / \max_{\boldsymbol{\theta} \in \Omega} L(\boldsymbol{\theta};\mathbf{X}) \right\}, \tag{13}
$$

where $L(\boldsymbol{\theta};\mathbf{x}) = p(\mathbf{x};\boldsymbol{\theta})$ is the likelihood function. For some constant A, we can use a test with critical region $\Lambda \leq A$.

If we define q as the difference in dimensionality of H_0 and $H_0 \cup H_1$, then we have that under the null, the asymptotic distribution of the test statistic is distributed according to:

$$
-2 \ln \Lambda \sim \chi_q^2, \quad \text{for } N \to \infty.
$$

We would like to emphasise that the GLRT test compares the likelihoods of parameters either belonging to the whole parameter space Ω, or to its subset $\omega \in \Omega$ (Eq. 13). This nesting of parameter spaces will be the basis for defining nested hypotheses in Definition (4).

4.1.2 Generalised Likelihood Ratio Test for Testing Causality

Let us refer to the null hypothesis of non-causality as it was formed in the Eq. 11. The likelihood ratio test can be rewritten in terms of a difference of two marginal log-likelihoods

$\ln p(\mathbf{Y} \mid \mathbf{X}^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m}; \boldsymbol{\theta}_B, \mathcal{M}_B) = \ln p(\mathbf{Y} \mid \mathbf{Y}^{-l}, \mathbf{Z}^{-m}; \boldsymbol{\theta}_A, \mathcal{M}_A)$, and it leads to the definition of a causality test statistic $L_{X \to Y|Z}$, first proposed by Amblard et al. (2012a):

$$L_{X \to Y|Z} = \max_{\boldsymbol{\theta}_B} \ln p(\mathbf{Y} \mid \mathbf{X}^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m}; \boldsymbol{\theta}_B, \mathcal{M}_B) - \max_{\boldsymbol{\theta}_A} \ln p(\mathbf{Y} \mid \mathbf{Y}^{-l}, \mathbf{Z}^{-m}; \boldsymbol{\theta}_A, \mathcal{M}_A). \tag{14}$$

In this paper we assume additive Gaussian errors, which allows us to calculate the marginal likelihoods analytically. For the calculations please refer to the Appendix 1. The resulting distributions are:

$$p(\mathbf{Y} \mid \mathbf{Y}^{-l}, \mathbf{Z}^{-m}; \boldsymbol{\theta}_A, \mathcal{M}_A) = \mathcal{N}(\mathbf{Y}; \boldsymbol{\mu}_A, \mathbf{K}_A + \boldsymbol{\Sigma}^A)$$
$$p(\mathbf{Y} \mid \mathbf{X}^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m}; \boldsymbol{\theta}_B, \mathcal{M}_B) = \mathcal{N}(\mathbf{Y}; \boldsymbol{\mu}_B, \mathbf{K}_B + \boldsymbol{\Sigma}^B).$$

If we use the hat notation for MLE estimators of the hyperparameters of the mean and covariance functions, then the test statistic is given by:

$$\begin{aligned}\hat{L}_{X \to Y|Z} = &-\big(Vec(\mathbf{Y}) - Vec(\hat{\boldsymbol{\mu}}_B)\big)^{\mathrm{T}} \Big(\oplus_{t=1}^{T} \hat{\mathbf{K}}_{\mathbb{Q}_{B,t}} + \hat{\sigma}_B^2 \mathbf{I}_{Tp' \times Tp'}\Big)^{-1} \big(Vec(\mathbf{Y}) - Vec(\hat{\boldsymbol{\mu}}_B)\big) \\ &+\big(Vec(\mathbf{Y}) - Vec(\hat{\boldsymbol{\mu}}_A)\big)^{\mathrm{T}} \Big(\oplus_{t=1}^{T} \hat{\mathbf{K}}_{\mathbb{Q}_{A,t}} + \hat{\sigma}_A^2 \mathbf{I}_{Tp' \times Tp'}\Big)^{-1} \big(Vec(\mathbf{Y}) - Vec(\hat{\boldsymbol{\mu}}_A)\big) \\ &- \ln \Big|\oplus_{t=1}^{T} \hat{\mathbf{K}}_{\mathbb{Q}_{B,t}} + \hat{\sigma}_B^2 \mathbf{I}_{Tp' \times Tp'}\Big| + \ln \Big|\oplus_{t=1}^{T} \hat{\mathbf{K}}_{\mathbb{Q}_{A,t}} + \hat{\sigma}_A^2 \mathbf{I}_{Tp' \times Tp'}\Big|.\end{aligned} \tag{15}$$

In the Eq. 15 we present a general form of the test statistic for multivariate time series, and in the special case of a univariate time series $\mathbf{Y}$ this simplifies to a form from the Eq. 16. Distinguishing between the two definitions can also be seen as a distinction between joint causality and marginal causality.

$$\begin{aligned}\hat{L}_{X \to Y|Z} = &-(\mathbf{Y} - \hat{\boldsymbol{\mu}}_B)^{\mathrm{T}} \Big(\hat{\mathbf{K}}_B + \hat{\sigma}_B^2 \mathbf{I}\Big)^{-1} (\mathbf{Y} - \hat{\boldsymbol{\mu}}_B) - \ln \Big|\hat{\mathbf{K}}_B + \hat{\sigma}_B^2 \mathbf{I}\Big| \\ &+(\mathbf{Y} - \hat{\boldsymbol{\mu}}_A)^{\mathrm{T}} \Big(\hat{\mathbf{K}}_A + \hat{\sigma}_A^2 \mathbf{I}\Big)^{-1} (\mathbf{Y} - \hat{\boldsymbol{\mu}}_A) + \ln \Big|\hat{\mathbf{K}}_A + \hat{\sigma}_A^2 \mathbf{I}\Big|.\end{aligned} \tag{16}$$

Under certain regularity conditions, with the assumptions of conditional independence of $\mathbf{Y}_t \mid \mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}$ for all t, and with the assumption that models A and B are nested (see 4.1.3) we can treat $L_{X \to Y|Z}$ as a GLRT and use the asymptotic results:

$$H_0 : \qquad 2\hat{L}_{X \to Y|Z} \sim \chi_q^2 \quad \text{as } T \to \infty,$$

where q is the difference in dimensionality between the parameter space for $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$.

4.1.3 Nested Models

An essential concept in our testing procedures is that of nested models. Its importance arises from the fact that the Generalised Likelihood Ratio Test (GLRT) on nested hypotheses has known asymptotic distribution.

Definition 4 Nested models. Two models: $\mathcal{M}_A$ parametrised by $\boldsymbol{\theta}_A$ and $\mathcal{M}_B$ parametrised by $\boldsymbol{\theta}_B$ are said to be nested if it is possible to derive one from another by means of parametric restriction, see Clarke (2000)

Intuitively, we could say that model A is nested in model B if the input space of model A is embedded in input space of model B, but the Definition 4 is formulated in terms of embedding of the model parameter spaces, rather than embedding of the input spaces. Formulating our Gaussian Process models A and B in such a way that they are nested according to the above definition is not always possible. This is because for the above definition of nested models we require the mean and covariance function to have parameters that correspond to the dimensionality of the input space, or that correspond to the inclusion or not of the input X.

In practice, when we talk about nested models we consider mean and kernel functions allowing the nested model representation. The simplest example of how the mean and kernel functions can allow nested models are for linear mean and kernel functions. Define $\mu_t([X_{t-1}, Y_{t-1}, Z_{t-1}]) = a_1 X_{t-1} + a_2 Y_{t-1} + a_3 Z_{t-1}$, which under restriction $a_1 = 0$ will become equivalent to a mean $\mu_t([Y_{t-1}, Z_{t-1}]) = a_2 Y_{t-1} + a_3 Z_{t-1}$, defined on the parameter space $[Y_{t-1}, Z_{t-1}]$. Analogously, for the linear kernel:

$$k_{t,s}([X_{t-1}, Y_{t-1}, Z_{t-1}], [X_{s-1}, Y_{s-1}, Z_{s-1}]) = \begin{bmatrix} X_{t-1} \\ Y_{t-1} \\ Z_{t-1} \end{bmatrix} \begin{bmatrix} A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,1} & A_{3,2} & A_{3,3} \end{bmatrix} [X_{s-1}, Y_{s-1}, Z_{s-1}]$$

restriction $A_{1,1}, A_{1,2}, A_{1,3}, A_{2,1}, A_{2,2}, A_{2,3}, A_{3,1} = 0$ will make this kernel equivalent to a linear kernel defined on $[Y_{t-1}, Z_{t-1}]$ with parameters $A_{2,2}, A_{2,3}, A_{3,2}, A_{3,3}$.

A popular kernel function that does not allow nested models is squared exponential kernel:

$$k([X_{t-1}, Y_{t-1}, Z_{t-1}], [X_{s-1}, Y_{s-1}, Z_{s-1}]) = \sigma_f^2 \exp\left(-\frac{([X_{t-1}, Y_{t-1}, Z_{t-1}] - [X_{s-1}, Y_{s-1}, Z_{s-1}])^{\mathrm{T}} ([X_{t-1}, Y_{t-1}, Z_{t-1}] - [X_{s-1}, Y_{s-1}, Z_{s-1}])}{2l^2} \right),$$

which, however, can be extended to a representation under an ARD structure, which does have a form that allows for nested models (see Subsect. 3.1 and the Table 1), if the division by a scalar lengthscale parameter $2l^2$ is replaced by a multiplication by the following matrix of lengthscale parameters: $diag([l_X^2, l_Y^2, l_Z^2])$.

If the nested model representation is not practical, then GLRT test should not be used. There are several approaches for non-nested models: modified (centered) log-likelihood ratio procedure – Cox procedure, "comprehensive model approach", "encompassing procedure", Vuong closeness test: likelihood-ratio-based test for model selection using the Kullback-Leibler information criterion. We refer the reader to the following papers (and references therein): Vuong (1989); MacKinnon (1983); Pesaran and Weeks (2001); and Wilson (2015).

5 Synthetic Data Experiments to Assess Proposed Causality Testing Framework

In this section, we seek to study the behaviour of our proposed methodology for GP testing of statistical causality relationships. In order to motivate the causality studies in this paper, we consider three illustrative nonlinear time series models. They will serve as references that we will apply our causality testing framework to, throughout the synthetic studies undertaken in the results analysis for testing power, sensitivity, and robustness of our proposed causality testing framework.

In particular the classes of model we have chosen as illustrations of data generating processes for the time series that will form inputs to our testing framework characterise a range of general model structures which allow for assessment of linear and nonlinear causality structures in the trend or the volatility or both components of the resulting data generating models.

Example Time Series Model Class 1: Structural Trend Based Causality Consider an autoregressive nonlinear model class comprised of structures incorporating time series with linear and nonlinear polynomial causality in the trend, with Gaussian noise.

$$\begin{aligned} X_t &= a_X X_{t-1} + \epsilon_X & \epsilon_X &\sim \mathcal{N}\big(0, \sigma_Z^2\big), \\ Y_t &= a_Y Y_{t-1} + b_Y X_{t-1} + \epsilon_Y & \epsilon_Y &\sim \mathcal{N}\big(0, \sigma_Y^2\big), \\ Z_t &= a_Z Z_{t-1} + b_Z Y_{t-1}^q + \epsilon_Z & \epsilon_Z &\sim \mathcal{N}\big(0, \sigma_Z^2\big), \end{aligned} \tag{17}$$

The examples that we will use will assume $q = 2$, which means that in the mean this time series will have a nonlinear causality in the direction $Y \to Z$, aside from the linear causality $X \to Y$.

We will express the model from the Eq. 17 in the form of three GPs, as in the Eq. 18. When generating the data, as Eq. 20 show, we will use Matern covariance functions with degrees of freedom $\nu = 1.5$, we will also extend the model to allow causal relationship in covariance – relationships, that were not existing in the time series formulations from Eq. 17.

A formulation of the time series from the Eq. 17 explicitly as GPs can be done according to the following conditional distributions:

$$\begin{aligned} X_t &= f_X(X_{t-1}) & f_X &\sim \mathcal{GP}\big(\mu_{X,t}, k_{X,t,t'}\big) \\ Y_t &= f_Y([Y_{t-1}, X_{t-1}]) & f_Y &\sim \mathcal{GP}\big(\mu_{Y,t}, k_{Y,t,t'}\big) \\ Z_t &= f_Z([Z_{t-1}, Y_{t-1}]) & f_Z &\sim \mathcal{GP}\big(\mu_{Z,t}, k_{Z,t,t'}\big) \end{aligned} \tag{18}$$

where the mean functions are linear:

$$\begin{aligned} \mu_{X,t} &= \mu_{X,t}(X_{t-1}) = a_X X_{t-1} && \text{no causality} \\ \mu_{Y,t} &= \mu_{Y,t}([Y_{t-1}, X_{t-1}]) = a_Y Y_{t-1} + b_Y X_{t-1} && \text{linear causality} \\ \mu_{Z,t} &= \mu_{Z,t}([Z_{t-1}, Y_{t-1}]) = a_Z Z_{t-1} + b_Z Y_{t-1}^2 && \text{nonlinear causality} \end{aligned} \tag{19}$$

and covariance functions incorporate the noise which was already defined as a GP:

$$
\begin{aligned}
k_{X,t,t'} &= k_{X,t,t'}(X_{t-1}, X_{t'-1}) = k^{Matern}_{l_a,\sigma_f}(X_{t-1}, X_{t'-1}) + \sigma_n^2 \delta_{t,t'} \\
k_{Y,t,t'} &= k_{Y,t,t'}([Y_{t-1}, X_{t-1}], [Y_{t'-1}, X_{t'-1}]) = k^{Matern}_{l_a,l_b,\sigma_f}([Y_{t-1}, X_{t-1}], [Y_{t'-1}, X_{t'-1}]) + \sigma_n^2 \delta_{t,t'} \\
k_{Z,t,t'} &= k_{Z,t,t'}([Z_{t-1}, Y_{t-1}], [Z_{t'-1}, Y_{t'-1}]) = k^{Matern}_{l_a,l_b,\sigma_f}([Z_{t-1}, Y_{t-1}], [Z_{t'-1}, Y_{t'-1}]) + \sigma_n^2 \delta_{t,t'}
\end{aligned}
\tag{20}
$$

Note that the main causality structure has been encoded in the mean functions, but the way the covariance functions are formulated allows some causality in the covariance in the directions $X \to Y$ and $Y \to Z$.

Example Time Series Model Class 2: Structural Causality Incorporated in Volatility The second causality structure has similar autoregressive and causal components to the Structure 1, but the error terms depend on past values of the other time series (so no autoregression in the covariance) via nonlinear functions f_y,f_z:

$$
\begin{aligned}
X_t &= a_X X_{t-1} + \epsilon_x \\
Y_t &= a_Y Y_{t-1} + b_Y X_{t-1} + \epsilon_y^*; \\
Z_t &= a_Z Z_{t-1} + b_Z Y_{t-1}^q + \epsilon_z^*;
\end{aligned}
\tag{21}
$$

where

$$
\begin{aligned}
\epsilon_y^* &= f_y\big(X_{t-1}, Z_{t-1}\big)\epsilon_y = \big(g_y(t) + c_y X_{t-1}^p + d_y Z_{t-1}^r\big)^2 \epsilon_y \\
\epsilon_z^* &= f_z\big(X_{t-1}, Y_{t-1}\big)\epsilon_z = \big(g_z(t) + c_z X_{t-1}^p + d_z Y_{t-1}^r\big)^2 \epsilon_z
\end{aligned}
\tag{22}
$$

The formulation above is general and the noise terms ϵ_y, ϵ_z can depend explicitly on time via the functions $g_y(t)$ and $g_z(t)$. We use c_y, c_z, d_y, d_z, p, q to denote constants. For this time series to be expressed in terms of GP we will have exactly the same general GP structure as for the time series 1 in the Eq. 18, and exactly the same mean functions – the Eq. 19. To construct the kernels that will match the covariance structure, we use the properties that summations and multiplications of kernels yield new kernels, for example as follows:

$$
\begin{aligned}
k_{X,t,t'}([X_{t-1}, Y_{t-1}, Z_{t-1}], [X_{t'-1}, Y_{t'-1}, Z_{t'-1}]) &= \sigma_n^2 \delta_{t,t'} \\
k_{Y,t,t'}([X_{t-1}, Y_{t-1}, Z_{t-1}], [X_{t'-1}, Y_{t'-1}, Z_{t'-1}]) &= k^{ts2}_{g,p,r,c_Y,d_Y}([X_{t-1}, Z_{t-1}], [X_{t'-1}, Z_{t'-1}])\sigma_n^2 \delta_{t,t'} \\
k_{Z,t,t'}([X_{t-1}, Y_{t-1}, Z_{t-1}], [X_{t'-1}, Y_{t'-1}, Z_{t'-1}]) &= k^{ts2}_{g,p,r,c_Z,d_Z}([X_{t-1}, Y_{t-1}], [X_{t'-1}, Y_{t'-1}])\sigma_n^2 \delta_{t,t'}
\end{aligned}
\tag{23}
$$

where: $k^{ts2}_{g,p,r,c,d}([W_t, V_t], [W_{t'}, V_{t'}]) = (g + cW_t^p + dV_t^q)^2 (g + cW_{t'}^p + dV_{t'}^q)^2$ is a kernel with the functions $g_y(t), g_z(t)$ simplified to a constant g. The notation $[W_t, V_t]$ should be understood as either $[X_{t-1}, Z_{t-1}]$ or $[X_{t-1}, Y_{t-1}]$.

Example Time Series Model Class 3: Causality Features in Presence of Long Memory The third data structure is a long memory process: ARFIMA(p,d,q), for $d \in [0, 0.5)$, with causality structure encoded in the form of external regressors:

$$
X_t - a_X X_{t-1} = \epsilon_{x,t} \tag{24}
$$

$$(Y_t - a_Y Y_{t-1} - b_Y X_{t-1})(1-B)^d = \Theta_Y(B)\epsilon_{y,t} \tag{25}$$

$$(Z_t - a_Z Z_{t-1} - b_Z Y_{t-1}^q)(1-B)^d = \Theta_Z(B)\epsilon_{z,t}, \tag{26}$$

where B is a backshift operator, the autoregressive coefficients for the time series Y_t, Z_t include external regressors, the moving average coefficient according to characteristic polynomial: $\Theta(B) = 1 - \theta_1 B - ... - \theta_q B^q$, and the long memory operator has linear process series expansion given for $d \in (0, 0.5)$ as follows:

$$(1-B)^{-d} = \sum_{k=0}^{\infty} \frac{\Gamma(k+d)}{\Gamma(k+1)\Gamma(d)} B^k.$$

In this example, there is no natural way to trivially develop a GP representation, however, it does not preclude fitting a misspecified model in order to screen for causality structures that may be present. We can fit such a model to partial observations of this reference example. This poses an interesting example to study the effect of model misspecification on the ability to detect linear and nonlinear causality structures.

5.1 Synthetic Data Experiments

In this section, we provide results for a series of tests of performance focusing on three key attributes of the proposed causality inference framework: power, sensitivity to parameters and robustness to model misspecification or parameter estimation errors. We perform these analyses for each of the three case study models introduced. We begin with sensitivity and misspecification tests, which we follow with experiments on the power of the test for simple and compound tests.

The sensitivity analysis shows how the test reacts to varying the parameter values used to generate the time series data in Example model 1, Eqs. 18–20. Here, we know the exact model so that a simple test is performed, where we assess its power over the parameter space.

The model misspecification tests show how the test reacts to discrepancy between the parameter values used to generate the time series data and the parameters used in the test statistic. This is a structured form of compound test analysis, since in practical settings in general the parameters will be estimated from data and then used in a compound testing procedure, in which the test statistics is a function of the estimated parameters.

We begin with two simple illustrative examples showing how the values of the test statistics from Eq. 14 change for different data samples, and what values of the χ^2 cdf they obtain. Throughout, we will perform analysis relative to the level of significance for the test of 10%. The Fig. 1 illustrates a compound test with optimised parameters – showing the values of test statistics $L_{X\to Y}$ vs $L_{Y\to X}$ and the 1-p values, or the evaluations of the distribution $\chi_2^2(2L_{X\to Y})$ vs $\chi_2^2(2L_{Y\to X})$. The data has been generated from causality structure 1 with strong causal effect $X \to Y$, with each of the 50 data replicate time series samples being of length 500 sample points.

The interpretation of the Fig. 1 is the following. From the left plot we can see that the test statistics $L_{X\to Y}$ has values which are separated from and considerably larger than the test statistics $L_{Y\to X}$. This by itself is an indication that the causal effect $X \to Y$ should be stronger than $Y \to X$. From the plot of cdf evaluations we observe that all of the values

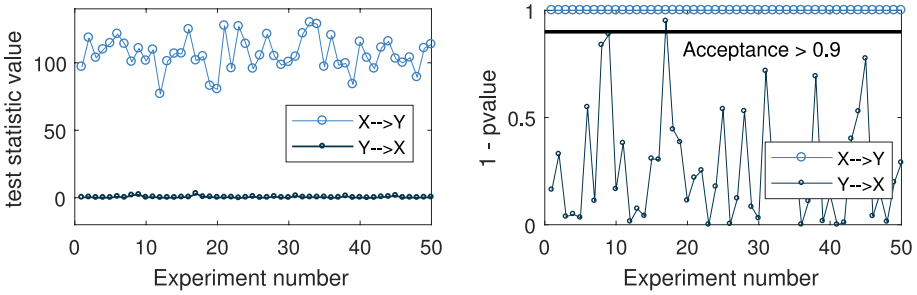

Fig. 1 Test statistics and corresponding cumulative density function evaluations. Causality structure 1, true parameters: $a_X = a_Y = a_Z = 0.3, b_Y = b_Z = 0.7, q = 2, l_a = l_b = e^{-6}, \sigma_f = e^{-10}, \sigma_n = 0.01$. The horizontal axis represents 50 separate trials, each with a time series of length 500

of $L_{X\to Y}$ are in the tail (with cdf values of exactly 1) and therefore the null hypothesis is strongly rejected at any confidence level, for each of the trials. This means that the estimator of the power of the test, i.e. the probability of rejecting the null hypothesis if it is not true, is very close to 1 for a very large range of confidence levels, certainly between 0.01% and 10%

This indicates that, as expected, the test performs very well in detecting the correct direction of causality – in this case $Y \to X$.

5.2 Model Sensitivity Analysis

It is important to ensure that, on one hand, the tests behave in a stable way when the parameters change – at least in some non-extreme region – and, on the other hand, that the tests are not heavily penalising misspecifications.

This test is performed for the first data structure, Eqs. 18–20. We use the following settings: Matern kernel, additive noise with variance of $\sigma_n^2 = 0.01$, grid of 21 different parameter values for each variation of the true model parameters assessed. For each experiment we consider 100 trials and the length of the simulated time series varies over range 20, 50, 100, 200, 500, 1000. We report rejection or lack of rejection of the test with the significance of $\alpha = 0.1$. The starting point is the parameter set: $a_X = a_Y = a_Z = 0.3$ and $b_Y = b_Z = 0.7$ (parameters of, respectively, autoregression and causality in the mean, as per Eq. 19), $l_a = l_b = e^{-1}, \sigma_f = e^{-3}, \sigma_n = 0.1$ (covariance parameters: autoregression, causality, multiplicative scaling, noise covariance, Eq. 20). Parameters are changed one at a time, and a new set of data is generated for each set of parameters.

We do not report results of the sensitivity test for the directions without causality: $Y \to X$ or $Z \to X$, as the test statistics in those cases will always be zero. When changing parameters in both models at the same time, we no longer use the true parameters, but we still compare models that are equivalent.

In the direction with causality $X \to Y$ we see that the behaviour of the test is very stable, with the changes in the frequency of rejection/non-rejection (here presented as estimated power of the test) influenced mostly by the sample size. The power of the test is the probability $P(H_0 \text{ rejected} | H_1 \text{ true})$, which in our case is estimated as $0.01 \cdot \sum_i^{100} F(2L_{X_i\to Y_i})$, where we have 100 trials, F denotes the cdf of χ_2^2 and 0.9 is 1 - confidence level.

Table 2 How power of the test changes with length of the time series (n) and changes of single parameters. Default parameters: $a_X = a_Y = a_Z = 0.3, b_Y = b_Z = 0.7, q = 2, l_a = l_b = e^{-1}, \sigma_f = e^{-3}, \sigma_n = 0.1$., one of the mean or covariance parameters changes ±50% in simulation and model as well. We look at time series of length $n = 20, 50, 100, 200, 500, 1000$. The parameter values correspond to the values in Fig. 2

XY	b_Y	a_Y	a_X	l_b	l_a	σ_f	σ_n^Y	σ_n^X
uniform?	+	+	+	+	+	+-	–	–
	min, max	min, max	min, max	min, max	min, max	min, max	min, max	min, max
$n = 20$	0.45, 0.98	0.84, 0.84	0.83, 0.88	0.84, 0.84	0.84, 0.84	0.80, 0.84	1.00, 0.09	0.40, 1.00
$n = 50$	0.76, 1.00	0.98, 0.98	0.97, 1.00	0.98, 0.98	0.98, 0.98	0.97, 0.92	1.00, 0.46	0.80, 1.00
$n = 100$	0.92, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 0.96	1.00, 0.70	0.91, 1.00
$n = 200$	0.99, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 0.87	0.99, 1.00
$n = 500$	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 0.99	1.00, 1.00
$n = 1000$	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00
YZ	b_Z	a_Z	a_Y	l_b	l_a	σ_f	σ_n^Z	σ_n^Y
uniform?	–	+-	+	+	+	–	–	–
	min, max	min, max	min, max	max, min	max, min	max, min	max, min	max, min
$n = 20$	0.02, 0.87	0.55, 0.22	0.33, 0.39	0.35, 0.35	0.35, 0.35	0.38, 0.63	0.35, 0.38	0.30, 0.96
$n = 50$	0.02, 1.00	0.72, 0.40	0.53, 0.61	0.55, 0.55	0.56, 0.56	0.79, 0.80	0.26, 0.77	0.50, 0.98
$n = 100$	0.05, 1.00	0.87, 0.52	0.69, 0.79	0.70, 0.71	0.71, 0.72	0.99, 0.85	0.21, 0.97	0.64, 1.00
$n = 200$	0.13, 1.00	0.98, 0.70	0.81, 0.93	0.86, 0.87	0.85, 0.89	1.00, 0.98	0.25, 1.00	0.77, 1.00
$n = 500$	0.31, 1.00	1.00, 0.90	0.94, 0.99	0.98, 0.98	0.97, 0.99	1.00, 1.00	0.80, 1.00	0.85, 1.00
$n = 1000$	0.50, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	1.00, 1.00	0.97, 1.00

Throughout these tests the 50% change in the parameters relates to the model parameters; The actual decrease/increase for covariance parameters is much bigger than for the mean, because the former is inputed to the algorithm as a logarithm

When compared to the $X \to Y$ direction, the results for $Y \to Z$ are less uniform, as shown in Table 2. The Table 2 demonstrates the power of the test for minimum and maximum of the parameter range, which is enough to portray the behaviour of the test for all parameters except σ_f for the $Y \to Z$ direction, for which local minimum can be seen in the Fig. 2. Based on the Table 2, and corresponding Fig. 2, we can also observe that the results for $Y \to Z$ are more sensitive to the change in parameters than the results for $X \to Y$, in particular the causal coefficient b_Z.

5.3 Model Misspecification Analysis

For the misclassification test we have chosen different starting settings for the covariance function $l_a = l_b = e^{-3}, \sigma_f = e^1$, which result in higher covariance, and much more pronounced effects of misclassification of covariance function parameters. Starting from the base set of parameters we alter one parameter at a time when calculating the test statistic; however, we use data generated for the base parameters: so that altered parameter is misspecified. It has to be emphasised that in the misspecification test a parameter will be altered for model A or model B, but not both.

Results of misclassification in the mean, which we do not report, are straightforward to understand and interpret. The power of the test depends mostly on the size of the

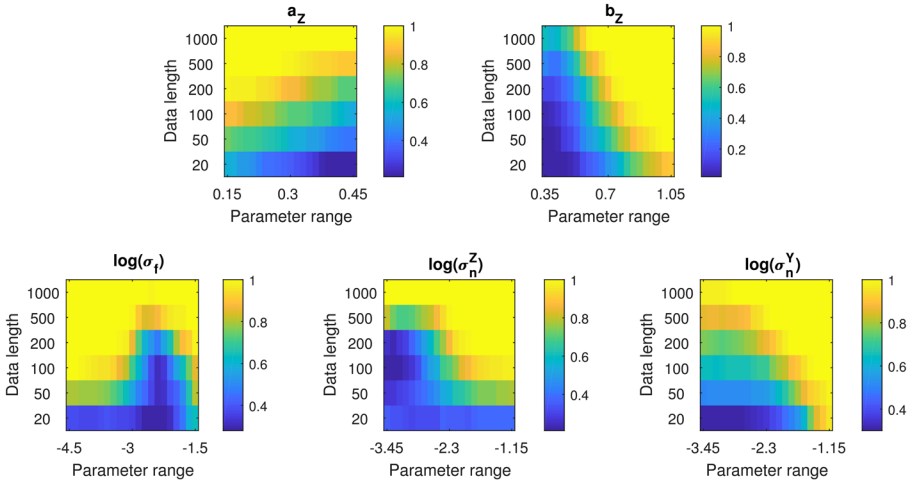

Fig. 2 Causality structure 1, direction $Y \to Z$ original parameters: $a_X = a_Y = a_Z = 0.3, b_Y = b_Z = 0.7$, $q = 2, l_a = l_b = e^{-1}, \sigma_f = e^{-3}, \sigma_n = 0.1$. Heatmaps show power of the test (hypothesis of no-causality rejected for cdf above 0.9) for different lengths of the time series and for one of the mean or covariance parameters changing $+-50\%$ in simulation and model as well

sample and, to a smaller degree, on the deviation from the true mean. For the direction where causality exists, the power of the test changes almost uniformly with the misclassification of the mean parameter. This is in line with observations that we will see repeatedly – that the power of the test is more robust to any parameter changes in the presence of causality in the mean.

Results of misclassification in the covariance, Figs. 3 and 4, are not so straightforward to understand and interpret. In particular, the performance of the tests seems to be more sensitive to the misclassification of the strength of the observation noise – this is not observed when parameters of the covariance (mainly σ_f) are smaller.

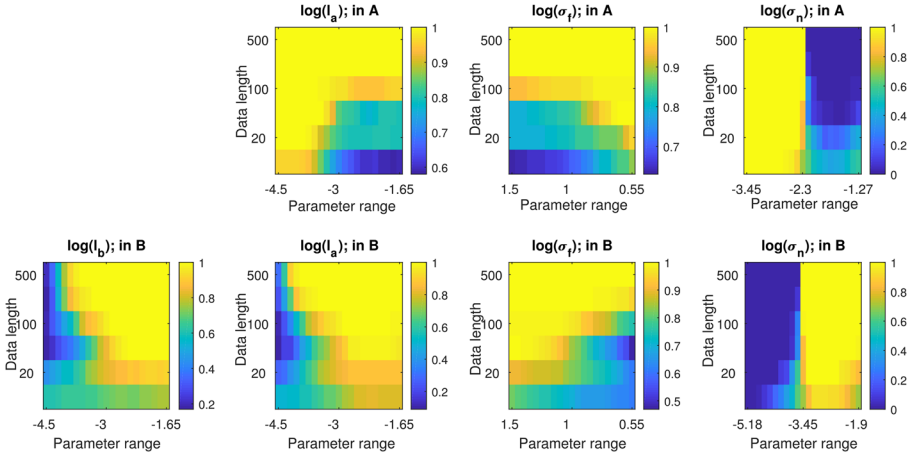

Fig. 3 Power of the test of the hypothesis of non-causality in the direction $X \to Y$ changes with the sample size and misspecification of a single hyperparameter (here – covariance parameters)

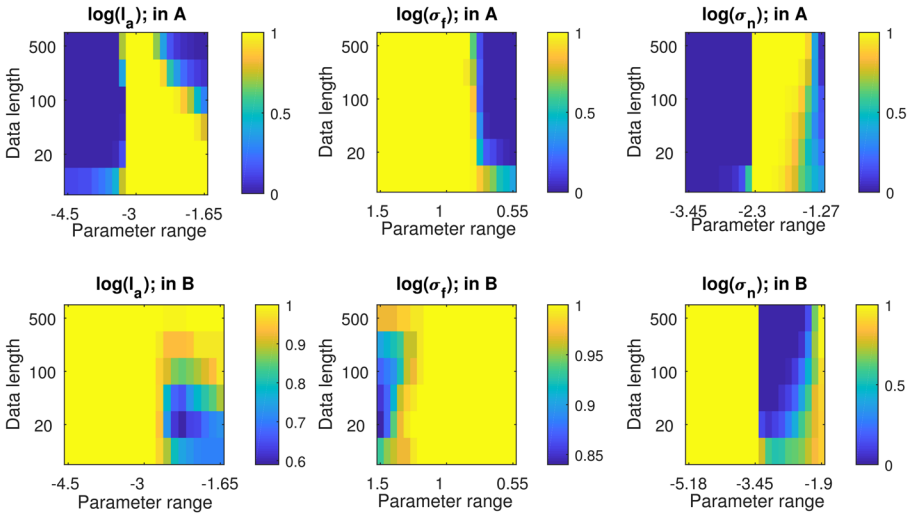

Fig. 4 How 1-rejection rate of the hypothesis of non-causality in the direction $Y \to X$ changes with the sample size and misspecification of a single hyperparameter (here – covariance parameters)

5.4 Power of the Hypothesis Tests: Simple Tests

Summary of the Section: *Analysing power of the test (1-rate of type II error) is a popular technique of assessing the quality of a test or a testing procedure. It is expected that the power of the test will increase with increasing sample size, and showing that this is indeed the case for our testing procedure will be the focus of this and the following sections. We start by analysing the results of simple tests, where exact parameters are used, and there is no effect of parameter misspecification. Strictly speaking, the simple test can be performed only for the first two data structures, as the third has been defined as an econometric model*

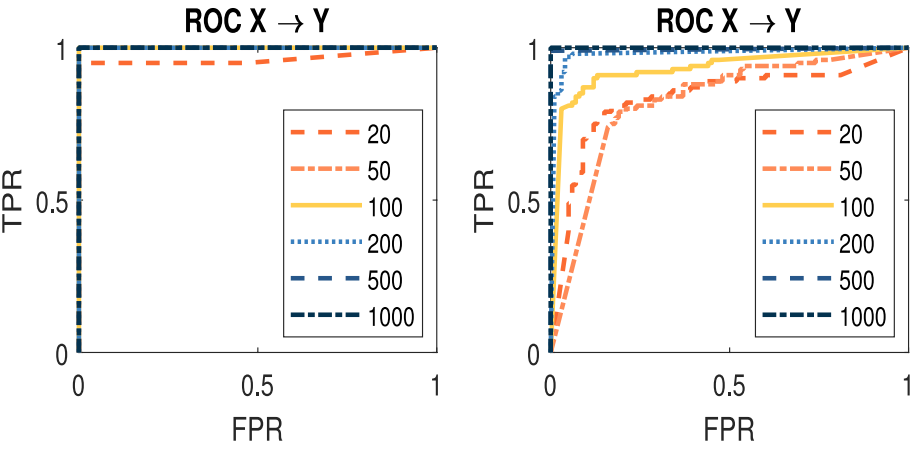

Fig. 5 Examples of parameter combinations for which the ROC curve shows different behaviour with longer sample (time series). True parameters: $a_X = 0.3, b_Y = 0.7$ in all 3 charts, the kernel parameters respectively: (left) $l_a = e^{-3}, l_b = e^{-1}, \sigma_f^2 = e^{-10}$, (right) $l_a = e^{-3}, l_b = e^{-1}, \sigma_f^2 = e^{-2}$

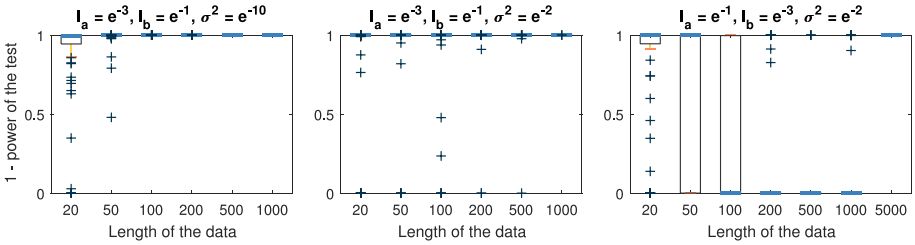

Fig. 6 Examples of parameter combinations that lead to different evolution of the test statistics distribution. True parameters: $a_X = 0.3, b_Y = 0.7$ in all 3 charts, the kernel parameters respectively: (left) $l_a = e^{-3}, l_b = e^{-1}, \sigma_f^2 = e^{-10}$, (middle) $l_a = e^{-3}, l_b = e^{-1}, \sigma_f^2 = e^{-2}$ and (right) $l_a = e^{-1}, l_b = e^{-3}, \sigma_f^2 = e^{-2}$. The right plot show an extreme case of performance decreasing with sample size for the typical range of sizes, hence the addition of results for data f length 5000

with no GP representation. However, for the third data structure, we perform a few tests with chosen parameters – to show the reaction of the test to certain properties of the data.

Example Time Series Model Structure 1 When using the exact parameters, as in a simple test, typically the behaviour for the Model 1 (Eqs. 19 and 20) is as expected: the power of the test increases with the sample size, and even in case of short time series the classification rule works well. This typical behaviour is illustrated in the left chart of Fig. 5 and in the left chart of Fig. 6. Figure 5 shows evolution of receiver operating characteristic (ROC) curves with increasing sample length, for two sets of parameters. When performing simple test, for most of the parameters, the ROC curves will show that positives and negatives are almost always properly classified, even for short time series – as seen in the example in the left chart of Fig. 5. This example represents testing of model 1 with true parameters: $a_X = 0.3, b_Y = 0.7$ for mean function, and $l_a = e^{-3}, l_b = e^{-1}, \sigma_f^2 = e^{-10}$ as the kernel parameters. The corresponding distributions of 1-power of the test can be seen on the left chart of the Fig. 6, in the form of boxplots. These distributions have medians at 1 for samples of length from 50 up, and no outliers for samples of length 500 and 1000.

The notable exceptions observed are as detailed below. Firstly, we show an example for which a higher rate of misclassification is seen, albeit it still decreases with the size of the sample. The right chart in the Fig. 5, has larger value of $\sigma_f^2 = e^{-2} \simeq 0.1353$, but with other mean and kernel hyperparameters remaining the same. The middle chart of the Fig. 6 shows that in this case, even for the sample of length 500 we still can observe some outliers with 1-power of the test at 0.

The right chart of the Fig. 6 shows an extreme case, where the power of the test degrades with length of the time series to a random coin flip on the hypothesis, although it improves if we consider exceptionally long samples of 5000 data points. We can see that the medians of the distributions of 1-power of the test drops from 1 to 0 for samples of length 100–1000, and gets back to 1 for sample of length 5000. The kernel hyperparameters in this case are equal: $l_a = e^{-1}, l_b = e^{-3}, \sigma_f^2 = e^{-2}$. This means signal variance at the same level as the less extreme case, but bigger autoregressive hyperparameter and smaller causal hyperparameter in the covariance function. Those parameter values, where increasing the sample size temporarily causes decrease of power of the test, can correspond to the dark areas from the Figs. 3 and 4.

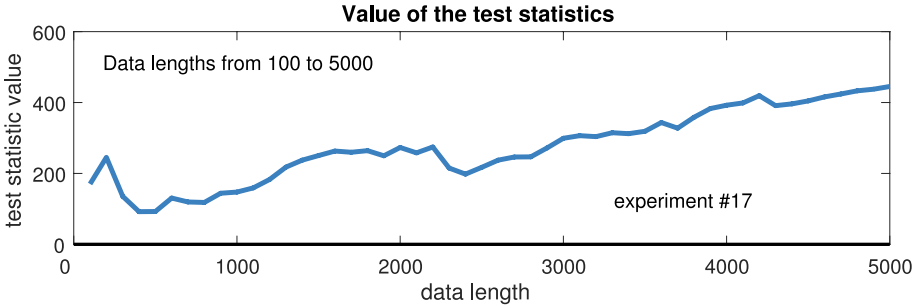

Fig. 7 Evolution of $L_{X\to Y}$ when (overlapping) data of different length is used. True parameters: $a_X = 0.3, b_Y = 0.7, l_a = e^{-3}, l_b = e^{-3}, \sigma_f^2 = e^{-2}$

The parameters that cause such behaviour is primarily the signal variance σ_f^2, and to a smaller extent l_a – the coefficient of autoregression in covariance function. The hyperparameter σ_f^2 increases the value of the covariance proportionately, while l_a - inversely and less than proportionately. Higher values of the covariance function mean higher volatility clustering, an effect which could compete with causality, but that could be less visible in short time series. We will not elaborate on this point here, but additional dependence structure can complicate the explanation of causality structure. Therefore longer time series appears necessary to correctly recognise causality in this case. The Fig. 7 shows the effect of length of a time series on the value of the test statistics $L_{X\to Y}$ for a particular combination of parameters. A single data set of length 5000 has been simulated and subsequently tests statistics have been calculated on the first 100, 200, 300, ...5000 data points. The chosen data set has a general trend of test statistics increasing for longer data lengths (as for all other data sets generated with the same parameters) but it shows to major dips of test statistics temporarily worsening.

The causal effect in the covariance function is difficult to observe. This is because on one hand, it seems to have a much subtler effect than the causality in mean, but also because it is entwined with other effects that can be observed for different parameter combinations. Figure 8 shows that for following parameters $b_Y = 0, a_Y = 0, l_a = e, l_b = e, \sigma_f^2 = e^4$ the causality in covariance is unambiguously observed already for sample size of 50. As a reminder, according to the Eq. 19, $b_Y = 0$ means no causality in the mean and $a_Y = 0$ means no autoregression in the mean.

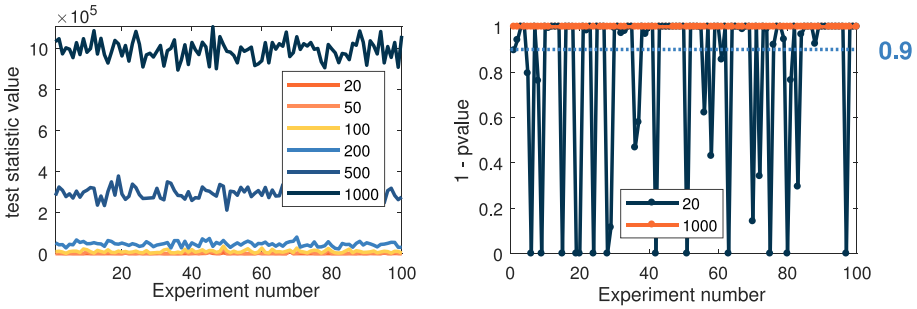

Fig. 8 Test statistics and the distribution evaluation: no causality in mean ($b_Y = 0$), no autocorrelation in mean ($a_Y = 0$), very large covariance parameters $l_a = e, l_b = e, \sigma_f^2 = e^4$. The right subplot does not explicitly show distribution evaluations for sample sizes from 50 to 500, because they are all equal 1 (just like for sample size 1000)

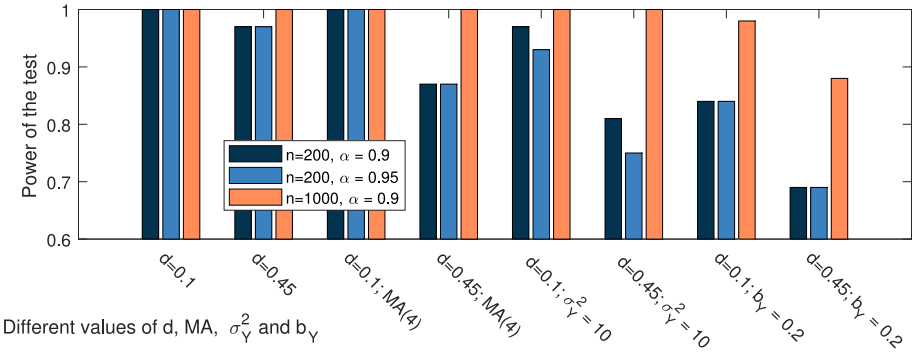

Fig. 9 The effect of the rate of decay of autocorrelation on the power of the test in model 3 varies strongly with different parameters

Example Time Series Model Structure 2 The results for testing of model 2, Eqs. 21–23, are just commented on here, since in the simple testing framework they do not show anything unexpected. In particular, the power of the test does increase with increasing length of the time series. Arguably, there is much less opportunity for problematic behaviour. This is firstly because the range of parameters which are available for the Example structure 2 is much narrower than for the Example structure 1 (i.e. parameters for which the series does not explode to infinity). Secondly, we assumed $cov(\epsilon_{Y_t}, \epsilon_{Y_{t'}}) = 0$, but if we did not we could have had again the problem with volatility clustering masquerading as causality.

Example Time Series Model Structure 3 We do, however, report a few observations on the testing of model 3. Firstly, model 3 does not have a GP representation, so when reporting on the results of the "simple test" in this case we do not perform a test with "true" parameters, but a test with fixed, rather than optimised, parameters. These observations become particularly interesting when compared with the results of the compound test for the data generated from the model 3. The main property of interest in the model 3 is the long memory, and this is what we concentrate on here. When analysing results for the data generated from model 3 (simple or compound test), on one hand, we expect that existence of the long memory will make recognition of causality more difficult, but on the other hand, we would like to see that causality can still be reasonably detected. Figure 9 shows how the power of the test is affected by increasing the long memory (values of the parameter $d = 0.1$ vs $d = 0.45$), and how this effect can be increased by changing other parameters (the degree of moving average from MA(1) to MA(4), noise covariance from $\sigma^2 = 0.1$ to $\sigma^2 = 10$, strength of linear causality from $b_Y = 0.7$ to $b_Y = 0.2$). It is worth emphasizing that decreasing strength of causality has the biggest influence, and is the only factor that affects the power of the test for long time series (length = 1000).

5.5 Power of the Hypothesis Tests: Compound Tests

Summary of the Section *Compound tests are two stage tests where both the likelihood as well as the model parameters are estimated. Robust estimation of parameters while possibly costly, is one of the most important pillars of robust testing with compound tests. In this section we want to draw attention of the reader to a few important phenomena: firstly, that the framework is much better*

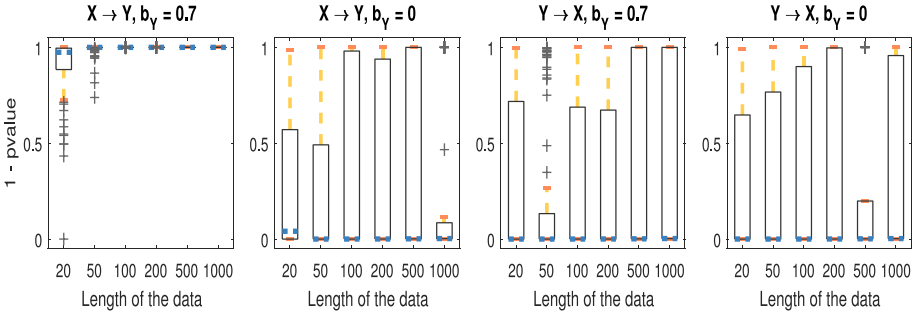

Fig. 10 Boxplots showing how the sample size affects distributions of the test statistics, in the case of existing causal effect (first subplot $X \to Y$ and $b_Y = 0.7$) and in the case where causal effect disappears due to causal coefficient equal to zero (second subplot $X \to Y$ and $b_Y = 0$), construction (third subplot $Y \to X$) or both (fourth subplot)

in picking up causality than accepting the lack of causality; and secondly, that even with strong model misspecification – which we will see for the model 3 – it is possible to identify causality.

One of the biggest factors influencing quality of the compound test is the efficiency of the optimisation algorithm. The objective function obtained from maximisation of the likelihood for parameter estimation produces generally a non-convex optimisation problem, which means that existence of local optima is likely. Using multiple starting points is highly recommended, but can potentially make the calculations very time consuming (our implementation involves a random grid of starting points). Using GPs with the assumptions we made in this paper (mainly: additive Gaussian noise) offers the advantage of being able to calculate the likelihood analytically. However, it is still possible that the data set can be so large, that this calculation will be prohibitively expensive. A popular approach in the literature is to decrease the dimensionality of the input data, see Snelson and Ghahramani (2007), or strive for efficient implementation, see Rasmussen and Williams (2006). An interesting and little known approach is to choose covariance function that promotes sparsity of the covariance matrix, as proposed by Melkumyan and Ramos (2009). Ensuring an approach is applicable to time series potentially adds a level of complication.

Example Time Series Model Structure 1 An observation that arguably holds for all data – not only the Model Structure 1 – is that when causality does exist in the data, the distribution of the test statistics estimator is much narrower than when there is no causality. An example is shown in the Fig. 10: the first plot shows that the causal signal can be picked up even for the shortest data, and the distribution of the tests statistics converges to value 1 already for length 100. When causality is not present (subplots 2 to 4) even for the longest used samples the distributions of test statistics are wide with median at zero, but 75th percentile often reaching close to 1.

Example Time Series Model Structure 2 The results of testing of model 2 show some very interesting behaviours. When fitting the model, we introduced model misspecification, because we allowed the structures to be the same for both directions. The first misspecification is in using polynomial means of second degree for $Y \to Z \mid X$ as well as $Z \to Y \mid X$. The second misspecification is in using the same volatility structure for both $X \to Y \mid Z$ and $Y \to X \mid Z$. As a result the estimated parameters in mean are often correctly estimated to be near zero, but the parameters in variance are strongly misspecified. The results still have reasonable power

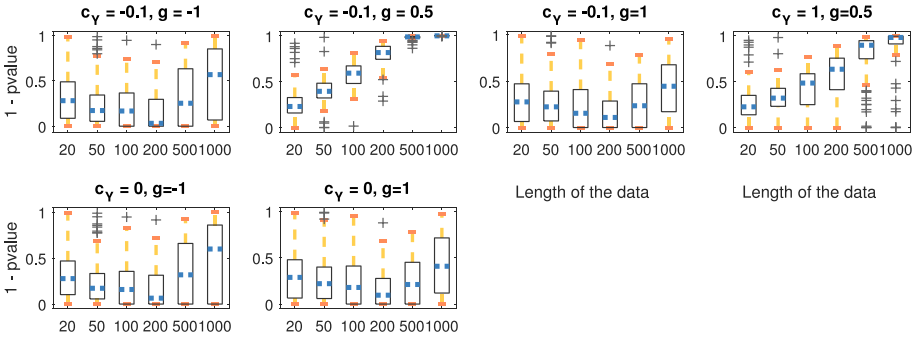

Fig. 11 Model 2, $X \to Y \mid Z$. Changes in recognition of causality with increasing sample size: different parameter settings. The top row shows the parameter settings where causal effect in covariance can be expected ($c_Y \neq 0$), while the bottom row shows cases where causality in covariance is not expected ($c_Y = 0$). In all the cases there was no causality in the mean ($b_Y = 0$)

of the test: the existence of causality is always correctly identified, however, in some cases the results could be interpreted as spurious causality. Also, like with model 1, there are cases where we seem to be spotting the causal effect in the covariance function when there is no causality in the mean, shown in the Fig. 11.

At the same time, we see that spurious causality signals are detected for the opposite direction: $Y \to X \mid Z$. Figure 12 shows how in the presence of causality $X \to Y \mid Z$ ($b_Y = 0.7$), the opposite direction also starts displaying causality with growing sample size. Explaining spurious causality is often complicated. In this case, we want to emphasise the following observations. First of all, the value of the test statistics is much bigger for the side where true causality exists, and a much smaller sample is needed to start indicating that causality with high confidence. Secondly, we run a misspecified model for the $Y \to X \mid Z$ direction (the misspecification is in the covariance function, with the multiplicative parameter σ_f having to equal zero to achieve properly specified function consisting of the multiplicative noise only), and even with multiple starting points, the optimised parameters are not as close to the true parameters as would be desired.

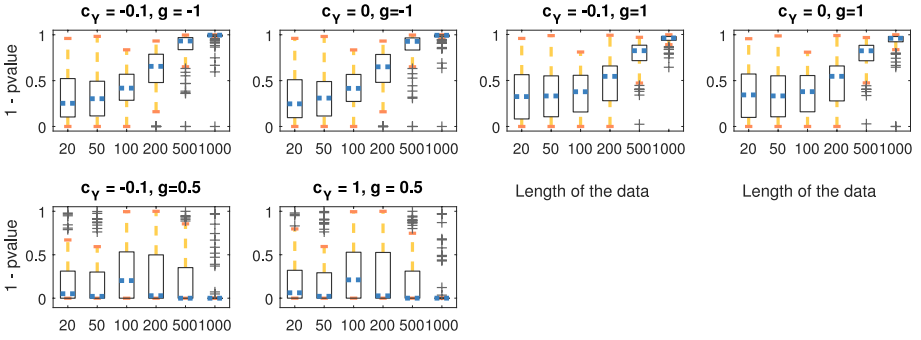

Fig. 12 Model 2, $Y \to X \mid Z$. Changes in recognition of causality with increasing sample size: different parameter settings. No true causality in the direction $Y \to X \mid Z$, but there was causality in the opposite direction ($b_Y = 0.7$). The parameter that affects recognition of spurious causality is the additive parameter g, whose higher absolute values tend to increase covariance

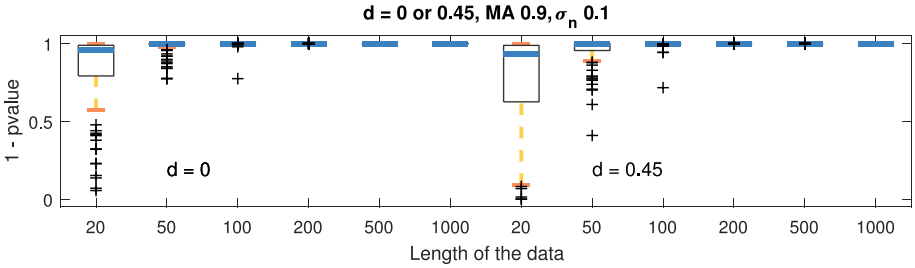

Fig. 13 Long memory barely affects the distribution of test statistics. This figure shows the distribution for the test statistics for $X \to Y$ for increasing length of the time series, first with no long memory $d = 0$, then with strong long memory $d = 0.45$

Example Time Series Model Structure 3 The results for the third data set exhibit a similar trend in the aspect that when a strong causal signal is present, it is correctly recognised. In case of lack of causality, or with very weak causal component, the distribution of the test statistics can be wide, but no spurious causality is detected. The data generated from model 3 has a long memory component, controlled by the parameter $d \in [0, 0.5)$, and one of the most interesting aspects is understanding the effect of long memory.

First of all, with the standard parameters, long memory hardly influences recognition of causality. Here, standard parameters are: strong causal component present ($b_Y = 0.7, b_Z = 0.7$), and the noise variance is not substantial ($\sigma_n^2 = 0.01$).

Figure 13 shows the distribution of test statistics when long memory is not present ($d = 0$), and when the effect of long memory is strong ($d = 0.45$), for different data lengths. The effect of changing parameters on the data generated from the model 3, in particular of changing the memory parameter d, is not significant. This seems unexpected at first, compared to the results of the simple test.

The explanation, however, lies in how the parameter estimation works, illustrated in the Fig. 14. The model is strongly misspecified and several properties of the data

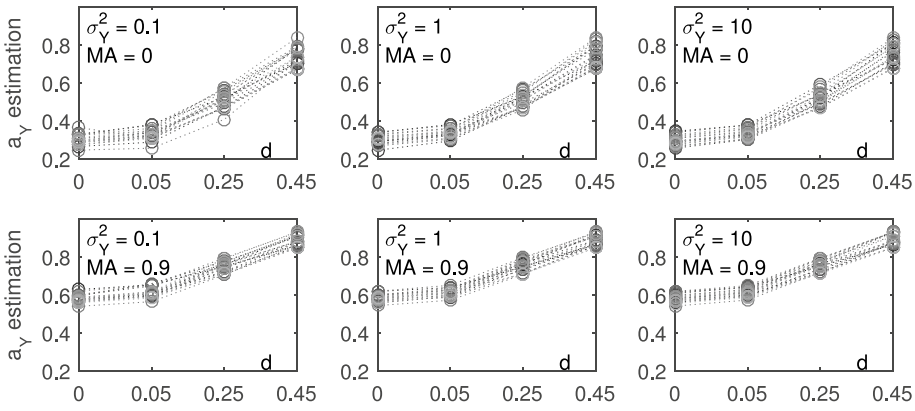

Fig. 14 How estimation of the autoregressive a_Y parameter "compensates" long memory or moving average effects. This figure shows the estimates of $\hat{a}_Y$ for different values of d, MA, σ_Y^2 and for different experiments, all of length 1000. It can be seen that the estimates strongly increase with increasing d and MA, and that this pattern appears for all values of the noise variance

are not well described by the model. Let us remember, though, that the long memory component has an infinite sum moving average representation, and the moving average model has an autoregressive representation. So the primary effect of increasing moving average part and the long memory part is the increase of parameters responsible for autoregression.

5.6 Comparison to Other Models

This section is provided to substantiate some of the claims we make about how our methods compare to existing methods. We provide three case studies, two of them compare our method to benchmark methods for causality: Granger causality and transfer entropy. The third case study compares our method to using generalised likelihood ratio test on a well specified econometric model (ARFIMA, example time series model class 3, Eqs. 24–26). What we show in our experiments is that our model achieves good results for all types of data, but in all cases, except for applying linear Granger causality test to linear causality, our method has superior asymptotic properties, as it reaches good power of the test for small samples.

Please note that in these case studies we concentrate on the ability to detect causality, and not on the time complexity of the algorithm.

Case Study 1: Granger Causality Granger causality can be seen as the original, but also the simplest method of assessing statistical causality. For Gaussian noise and linear causal relationship, Granger causality is arguably the best method, given that the test statistics have known asymptotic distributions, and estimators have excellent numerical properties. What is more, Granger causality can perform well for a range of data that departs from the model assumptions.

In this, and in the next case study, we will use four data sets, designed to show the effect of the departure from the assumption of data with linear dependence, stationary distributions, and Gaussian noise (as introduced earlier in the Eq. 17), replicated below with slight modifications:

$$\begin{aligned} X_t &= a_X X_{t-1} + \epsilon_X, \\ Y_t &= a_Y Y_{t-1} + b_Y X_{t-1} + \epsilon_Y, \\ Z_t &= a_Z Z_{t-1} + b_Z Y_{t-1}^2 + \epsilon_Z, \quad \epsilon_X, \epsilon_Y, \epsilon_Z \sim i.i.d \text{ white noise}, \end{aligned} \tag{27}$$

The data model from Eq. 27 exhibits two causal relationships. The causal relationship $X \to Y$ is – if we assume Gaussian white noise – of the type that Granger causality has been designed to model: linear, stationary, with Gaussian distributions. We will call this a base case (set one), and we will consider three other cases, each presenting a departure from one of those three properties. The causal relationship $Y \to Z$ is not linear, and it forms the set 2. We will also consider what happens to the ability to detect relationship $X \to Y$, if we changed Gaussian noise to t-student noise (set 3), and if we changed stationary to non-stationary marginal distributions (set 4; in this case we use polynomial covariance, please refer to the Table 1). These four set and their properties are summarised in the Table 3.

We present the results for the Granger causality method, using the GCCA toolbox. The test statistic used in the toolbox is the measure of linear feedback introduced by Geweke (1982), as in the Eq. 6. The corresponding test used for testing the null

Table 3 Data used for Case Study 1 and 2. Causal relationship number 1 is the base case: linear, with stationary marginal distributions and Gaussian noise. The three other causal relationships show three types of departure from the base case

set nr.	1	2	3	4
direction	$X \to Y$	$Y \to Z$	$X \to Y$	$X \to Y$
linearity	linear	nonlinear (square)	linear	linear
noise	Gaussian	Gaussian	t-student, 5 df	Gaussian
stationarity	stationary	stationary	stationary	non-stationary

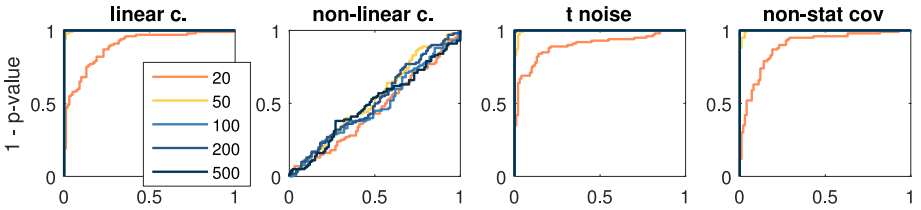

Fig. 15 ROC curves for the data sets 1-4 from the table, calculated with (linear) Granger causality, tested with the GCCA toolbox

hypothesis of lack of causality is the F-test. The results are presented graphically in the Figs. 15 and 16.

The results of using Granger causality can be summarised by two main observations. Firstly, for strong linear causality relationship, the linear Granger causality test is very robust and practical even if we do not observe Gaussian noise or stationary covariance. Secondly, for nonlinear causality, the linear Granger causality method behaves no better than a random guess, regardless of the data size. How does that compare to our method? The Fig. 16 shows that for strong, linear causality, our method is not as robust as linear Granger causality, and requires a bigger sample. However, our method can successfully detect nonlinear causality. For the data with t-distributed noise, we present results for the test statistic calculated by assuming the correctly specified model, and using an approximate method[3].

Case Study 2: Transfer Entropy We have used the same data structures as described in the Eqs. 27 together with the Table 3. The results are graphically shown in the Fig. 17.

Transfer entropy is a popular method used as a nonlinear extension of the linear Granger causality (for Gaussian distributions these two methods are equivalent). It is able to consider wider range of data types and relationships, however it is much more difficult to estimate. Compared to our method, transfer entropy requires much larger data samples, and at the same time it is not able to deal with model structures like long memory, non-stationarity, etc. Comparing Figs. 15 and 17 shows inferior performance of transfer entropy to our method in each of the four cases, and inferior to (linear) Granger causality in three cases. Transfer entropy is better than Granger causality in recognising nonlinear causality,

[3] Assuming a misspecified model with Gaussian likelihood, and then using the exact method to optimise parameters brings comparable results in this case.

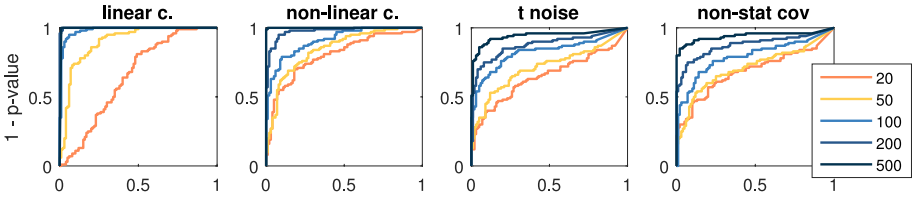

Fig. 16 ROC curves for the data sets 1-4 from the table, tested with our method

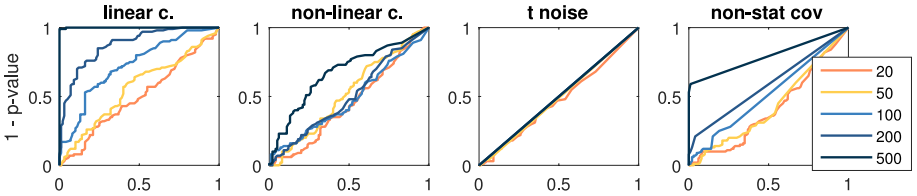

Fig. 17 ROC curves for the data sets 1-4 from the table, calculated with transfer entropy based on the binning algorithm

however, only for the sample of size 500 is transfer entropy performing recognisably better than a random choice.

What is not shown in the results, but for the sake of fairness needs to be mentioned, is the fact that transfer entropy is much faster than our method, with the current implementation.

Case Study 3: ARFIMA Model The data that was used for this example has been generated according to an ARFIMA (1,d,1) model with external regressors, Eqs. 24–26, can be represented in a form emphasising the autoregressive part (this is possible because we restricted the choice of d to (0, 0.5)):

$$
\begin{aligned}
X_t &= a_X X_{t-1} + \epsilon_X \\
Y_t &= a_Y Y_{t-1} + b_Y X_{t-1} + \epsilon^*_{y,t}, \quad \epsilon^*_{Y,t} = (1-B)^{-d}\Theta_Y(B)\epsilon_{Y,t} \\
Z_t &= a_Z Z_{t-1} + b_Z Y^q_{t-1} + \epsilon^*_{z,t}, \quad \epsilon^*_{Z,t} = (1-B)^{-d}\Theta_Z(B)\epsilon_{Z,t}.
\end{aligned}
$$

Table 4 Nine sets of parameters for the ARFIMA model, that were used in our analysis, in the Case Study 3

Set nr	a_Y	b_Y	MA	d	
1	0	0	0	0	pure noise
2	0.3	0	0	0	ARFIMA(1,0,0)
3	0.3	0.7	0	0	ARFIMA(1,0,0) and causality
4	0	0	0.9	0	ARFIMA(0,0,1)
5	0	0	0	0.49	ARFIMA(0,d,0)
6	0.3	0.7	0	0.25	ARFIMA(1,d,0) and causality
7	0.3	0.7	0.9	0	ARFIMA(1,0,1) and causality
8	0.3	0.7	0.9	0.25	ARFIMA(1,d,1) and causality
9	0.3	0.7	0.9	0.49	ARFIMA(1,d,1) and causality

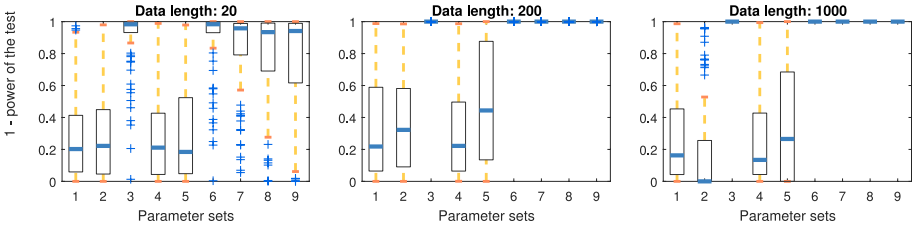

Fig. 18 Distributions of test statistic for GPC method, shown for three lengths of the time series, and for 9 data sets

We estimate data using modified MATLAB code ARFIMA-SIM by Fatichi (2009). For fitting the ARFIMA with external regressors we use the rugarch R library. We present results for nine parameter settings, which are listed in the Table 4 .

We present the results of using our causality method to estimate causality in Fig. 18, while the results of using a fully specified likelihood of the ARFIMA model are shown in the Fig. 19.

Our method is operating on the GP model representation, which is clearly misspecified. However, that does not prevent our model from detecting causality even for the smallest samples of length 20. That is not the case for using the well specified ARFIMA model and estimated likelihood – in this case a very large sample is needed for the estimation to even converge – data of length 1000 is required for the calculation of the results for all 9 data sets.

6 Real Data Experiments

In this section we apply the testing procedures to analyse commodity futures data.

In our analysis we use the following data: 1 and 36 month expiry oil futures contracts, obtained from futures curves built on the basis of West Texas Intermediate (WTI) Crude oil futures prices traded on the New York Mercantile Exchange, as described by Ames et al. (2016). The effect of the currency level, captured by the US Dollar Index DXY, is constructed as an index of USD relative to EUR, JPY, GBP, CAD, SEK, CHF. Thirdly, we also use a widely considered proxy for convenience yield based on a component related to transportation expense, given by the cost of freighting and short term

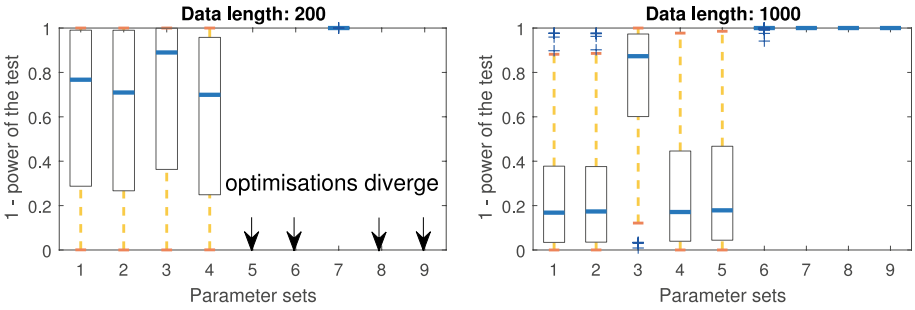

Fig. 19 Distributions of test statistic for the AFRIMA likelihood method, shown for two lengths of the time series, and for 9 data sets

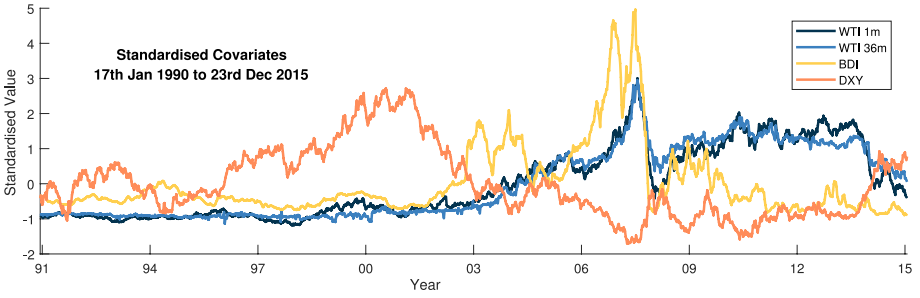

Fig. 20 1 and 36 month oil futures (WTI), Baltic Dry Index (BDI), Dollar index (DXY), all standardised

storage, measured by the Baltic Dry Index (BDI), see Ames et al. (2016). There is a stochastic functional relationship between commodity futures contracts of different maturities (term structure) based on: spot price, convenience yield, interest rate, and dollar value. Convenience yield is very hard to model, but can be captured to some extent by BDI, and the interest rate can be proxied by the time value of money expressed by the futures contracts. Hence the choice of both long and short dated futures contracts for our analysis. The Fig. 20 shows the four covariates from 17th Jan 1990 to 23rd Dec 2015. For literature studying classical relationships between these data, we refer to: Ames et al. (2016), Bakshi et al. (2010) and Dempster et al. (2012).

6.1 Interpreting Causal Relationships

The study performed here uses causality testing to demonstrate the risk factors that investors should consider in their decision process. It also shows how speculators in currency markets and futures markets have a propensity to respond to information observed at different lags and the time it takes them to re-adjust the expectations for futures market hedging or speculation in light of this information.

Figures 21, 22, 23, 24 present the changing significance of causal relationships between the dates 17th Jan 1990 to 23rd Dec 2015. The four pairs that we look at, and the abbreviations that we will use are as follows: 1 month oil futures (1m WTI) and freighting/ storage index (BDI), 36 months oil futures (36m WTI) and freighting/ storage index, 1 month oil futures and dollar index (DXY), 36 months oil futures and dollar index. We are presenting causal reactions at two lags: one week, which can be seen as nearly instantaneous, and eight weeks. Figures 21–24 show charts smoothed with cubic spline smoothing, which makes it easier to observe the main trends, in particular in the case of lags of 8 weeks.

Markets learn from the news and facilitate them into the price, according to the efficient market hypothesis[4], to which we subscribe (Fama (1970); Fama and French (1988); Campbell

[4] **Efficient market** can be defined as " [a market that] (...) *do not allow investors to earn above-average returns without accepting above-average risks.* (...) *Markets can be efficient in this sense even if they sometimes make errors in valuation*", Malkiel (2003). Market efficiency, as it is understood nowadays, is the belief that new information is reflected in price quickly and accurately, but not necessarily instantaneously. See Malkiel (2003) and sources therein.

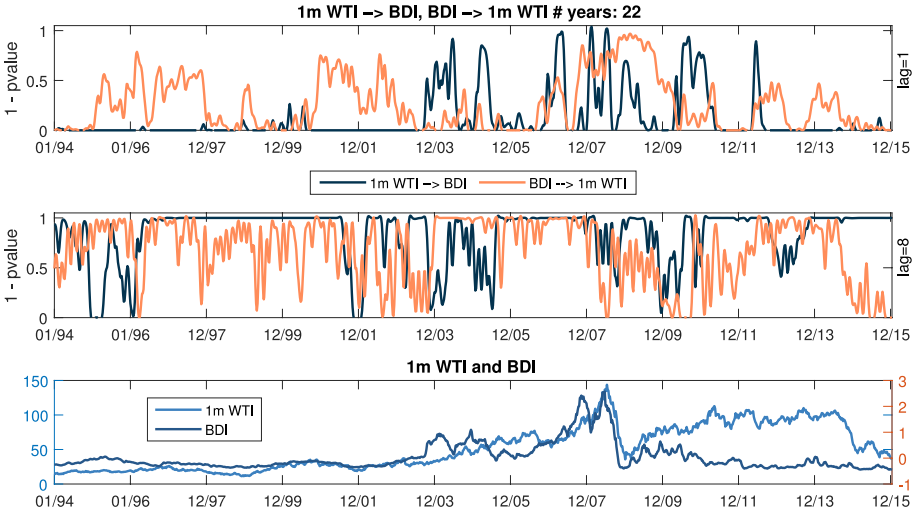

Fig. 21 Evolution of the causal influence: 1-pvalues of the test statistic for 1 months WTI and BDI, with 1 lags (top subplot) and 8 lags (middle subplot), rolling window of 104 weeks and cubic spline smoothing. Bottom subplot presents prices of 1 month oil futures contracts and historical values of BDI index

and Shiller (1988); Campbell et al. (1997); Malkiel (2003)). We want to learn which variables have effect on price formation, and at what time horizon. We also want to relate to the fact that the three different classes of investments (oil futures, currencies, physicals) have different investor profiles, and thus we expect a difference in the type and speed of reaction. The last question that interests us, is whether the results confirm the intuition that regimes affect the direction and significance of causal influence.

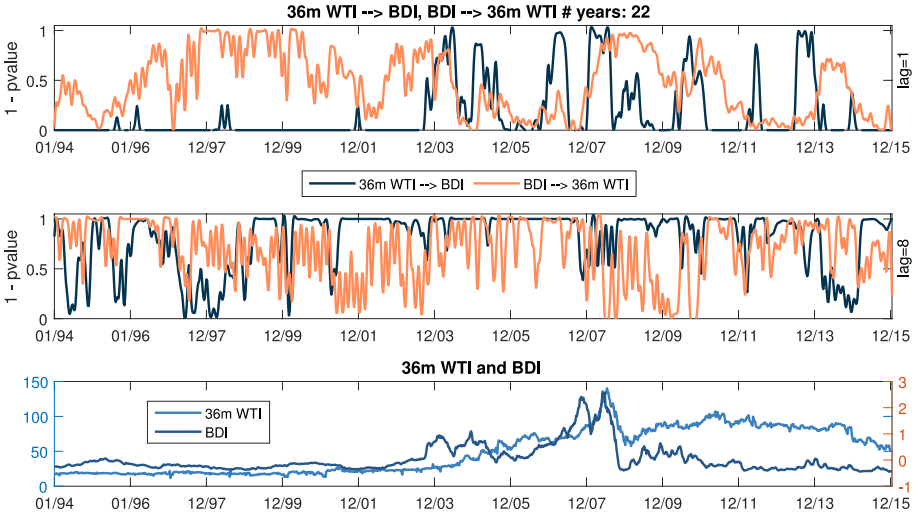

Fig. 22 Evolution of the causal influence: 1-pvalues of the test statistic for 36 months WTI and BDI, with 1 lags (top subplot) and 8 lags (middle subplot), rolling window of 104 weeks and cubic spline smoothing. Bottom subplot presents prices of 1 month oil futures contracts and historical values of BDI index

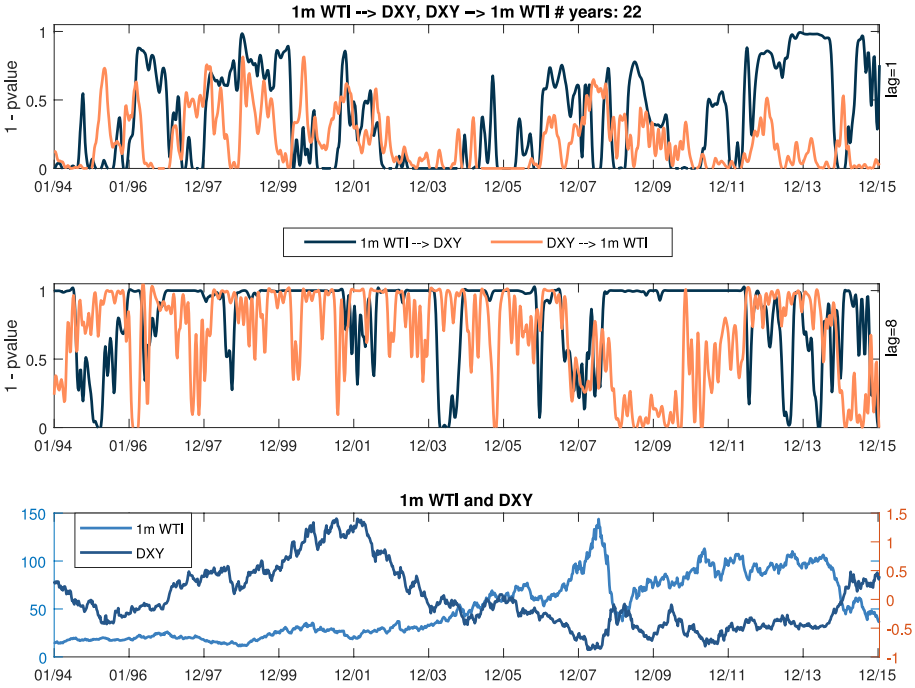

Fig. 23 Evolution of the causal influence: 1-pvalues of the test statistic for 1 months WTI and DXY, with 1 lags (top subplot) and 8 lags (middle subplot), rolling window of 104 weeks and cubic spline smoothing. Bottom subplot presents prices of 1 month oil futures contracts and historical values of dollar index

The interplay between WTI oil futures and the cost of freighting (BDI) Market participants investing in freighting are likely to be interested in the ownership of the physical asset, therefore BDI can be used as a proxy for convenience yield. It is expected that the WTI oil futures will not have instantaneous effect on the BDI, which is confirmed by our analysis showing that the causal direction from WTI to BDI is generally not statistically significant at 1 lag (Figs. 21 and 22, top subplots).

The effect to which the WTI futures incorporate the BDI movements varies across maturities. Short contracts have not been reacting to BDI changes in 1 week, with the exception of 2008/2009, which was a reaction to crisis. Similar response can be seen for longer maturities, however for longer maturities we observe the BDI→36m WTI to be significant through late nineties.

At 8 lags, we observe that the causal effects are significant in both directions, majority of the time. This can be seen as markets being able to absorb the information and adjust the expectation. For the times when this relationship breaks, investors use other sources, to inform their long term perception of risk and expectations: for example as a result of the 2008 crisis investors across many markets were decreasing their exposure to risk. In late nineties, as well as in 2014, we can observe a divergence of reactions of BDI to short and long term oil futures at 8 week lags: this could be seen as investors using outside information to decide on their long term expectations: for example about advancement in methodology or legislation pertaining renewable energy.

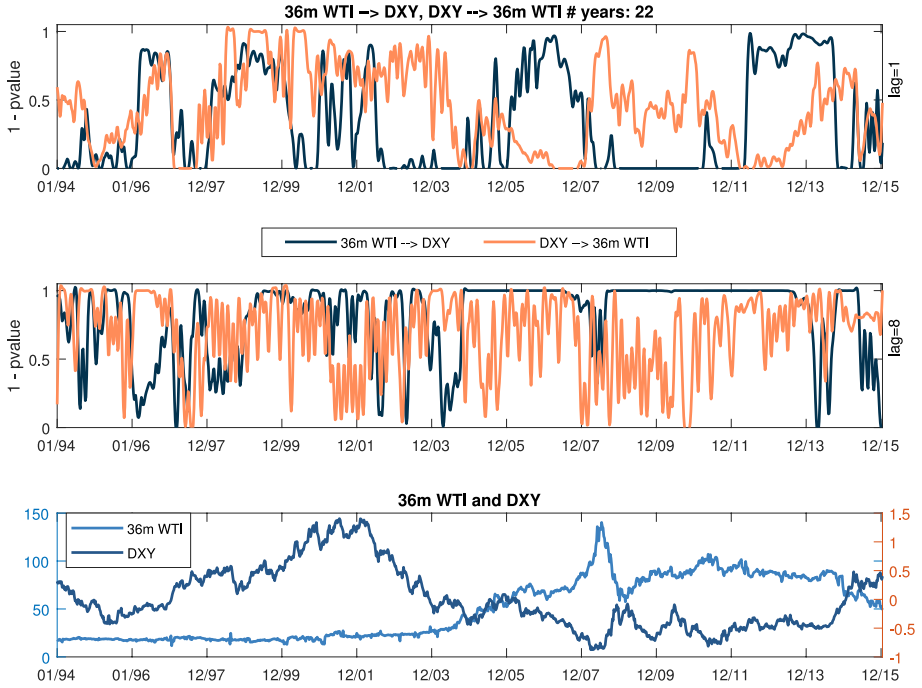

Fig. 24 Evolution of the causal influence: 1-pvalues of the test statistic for 36 months WTI and DXY, with 1 lags (top subplot) and 8 lags (middle subplot), rolling window of 104 weeks and cubic spline smoothing. Bottom subplot presents prices of 1 month oil futures contracts and historical values of dollar index

The interplay between WTI oil futures and the dollar index (DXY) The dollar index is a weighted geometric mean of the dollar's value relative to a basket of foreign currencies: Euro (EUR) 57.6% weight, Japanese yen (JPY) 13.6% weight, Pound sterling (GBP) 11.9% weight, Canadian dollar (CAD) 9.1% weight, Swedish krona (SEK) 4.2% weight, Swiss franc (CHF) 3.6% weight. The Canadian dollar is considered a commodity currency, while the Japanese yen is particularly sensitive to changes in oil prices due to Japan importing almost all of its oil. Therefore market expectations towards dollar index will incorporate to a large degree the expectations that arise from the oil market.

Following the results from the Fig. 24, there is evidence to suggest that DXY drives longer dated futures more strongly. At the same time, when comparing top charts from Figs. 22 and 24, we notice similarity in causal pattern between DXY → 36m WTI and BDI → 36m WTI, in particular during the nineties. This could suggest another direct or indirect factor, common for the two causal direction, for example general attitude to risk.

We look at Markov Switching Model, to analyse if DXY and BDI will have similar patterns of states for volatility, when explained with VIX. We use the following models:

$$D_t = \alpha_{1,S_t} + \alpha_2 V_t + \epsilon_t^D \quad \epsilon_t^D \sim \mathcal{N}(0, \sigma^2_{D,S_t}), \tag{28}$$

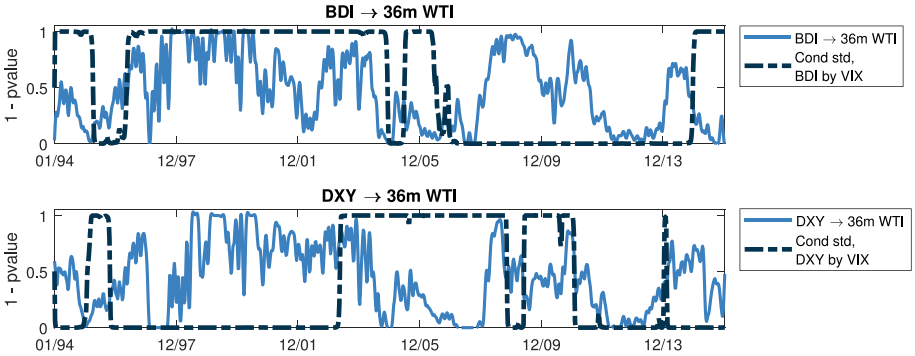

Fig. 25 Conditional standard deviation of error of the regime switching model explaining DXY or BDI with constant and VIX, scaled to $[0, 1]$, compared to the 1-pvalue of the BDI → 36m WTI and DXY → 36m WTI, for 1 lag

$$B_t = \beta_{1,S'_t} + \beta_2 V_t + \epsilon_t^B \quad \epsilon_t^B \sim \mathcal{N}(0, \sigma^2_{B,S'_t}), \tag{29}$$

where: S_t and S'_t, which we assume to only take values 1 and 2, are the states at time t for DXY and BDI respectively, $\sigma^2_{D,S_t}, \sigma^2_{B,S'_t}$ are the variances of the innovation at state S_t, S'_t, $\alpha_{1,S_t}, \beta_{1,S'_t}$ are the mean coefficients at state S_t, S'_t, and $\epsilon_t^D, \epsilon_t^B$ are innovations.

Figure 25 presents the conditional standard deviation of error term for regime switching models from Eq. 28 and 29, scaled for clarity to $[0, 1]$, and superimposed on the power of the tests of BDI → 36m WTI and DXY → 36m WTI, for 1 lag. First of all, for BDI it is the decreased conditional volatility that coincides with higher evidence of causality, while for DXY it is the increased volatility. However the persistence of high evidence for causality from 1996 to 2002 for both DXY → 36m WTI and BDI → 36m WTI, coincides with the persistence of one state for conditional standard deviation of respective covariates over that period of time. This suggests that the perception of market risk as seen via VIX is a common driving factor for during the nineties, a factor which can supersede other dependencies.

6.2 Influence of the Absolute Value of the Oil Prices On the Causal Structure

During the times when world oil prices are seen as high, it is more reasonable to expect investments in oil infrastructure as well as storage and transport. Therefore, we would expect that the absolute level of the oil price affects the behaviour (direction, strength, persistence) of causality. To test this, we compare the causal structure, as well as the fitted models, during the period of low prices: 17.01.1990 – 11.08.1999 (bellow $40), and period of high prices: 26.05.2004 – 11.03.2009 (above $90). We will be interested in the relative difference between the fitted mean values, as well as the relative difference between hyperparameters (coefficients of the mean): autoregressive and causal. For that we will be using two sample mean test. Please note, that while we are particularly interested in the change of regime in the fitted models, we also check the regime change of the causal test statistics – this is because we were earlier making a point of being able to detect causality even in misspecified models!

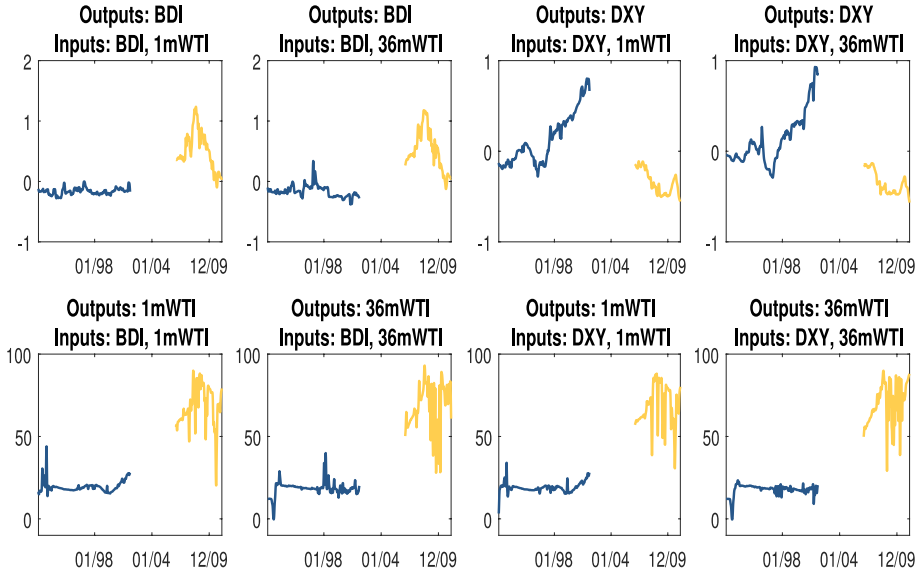

Fig. 26 Mean function estimations for each of the pairs of time series. The two colours represent the two different segments: 17.01.1990 – 11.08.1999 (oil prices bellow $40), and period of high prices: 26.05.2004 – 11.03.2009 (oil prices above $90)

Lets assume that for each of the pairs: 1 m WTI and BDI, 36 m WTI and BDI, 1 m WTI and DXY, 36 m WTI and DXY, we take X_t to denote one of the time series from the pair, and Y_t - the other:

$$\begin{aligned} X_t &= f_X([X_{t-1}, Y_{t-1}]) \quad f_X \sim \mathcal{GP}\big(\mu_{X,t}, k_{X,t,t'}\big) \\ Y_t &= f_Y([Y_{t-1}, X_{t-1}]) \quad f_Y \sim \mathcal{GP}\big(\mu_{Y,t}, k_{Y,t,t'}\big), \end{aligned}$$

with the usual notation. We denote M_t^X and M_t^Y as time series of values of the mean functions fitted by the models used for causality testing on rolling windows. Figure 26 shows the two segments of the fitted means: segment corresponding to prices below $40 and above $90, and in the Fig. 27 these have been additionally filtered according to the significance of the causal hypothesis. The mean function estimations are calculated on moving windows, with one mean function estimation equal to a mean of fitted values for the respective window.

For each of the pairs, we performed a two means test:

$$\begin{aligned} H_0 &: \quad mean(M^X_{01.90-08.99}) = mean(M^Y_{05.04-03.09}) \\ H_1 &: \quad mean(M^X_{01.90-08.99}) \neq mean(M^Y_{05.04-03.09}) \end{aligned}$$

We have run the popular student-t distribution two means test, as well as a two means test using sieve bootstrap to correct for serial dependence. The results are unanimously rejecting the hypotheses of equal means.

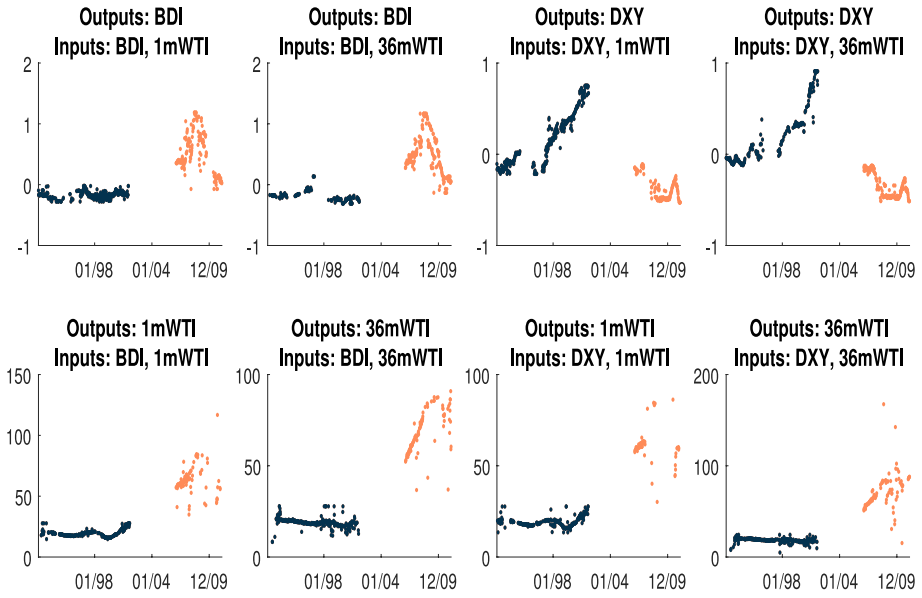

Fig. 27 Mean function estimations for each of the pairs of time series, shown only for the time points for which the hypothesis of lack of 8 week lag causality has been rejected at the level of $\alpha = 5\%$. The two colours represent the two different segments: 17.01.1990 – 11.08.1999 (oil prices bellow \$40), and period of high prices: 26.05.2004 – 11.03.2009 (oil prices above \$90)

6.3 Contrasting the Granger Causality With Our Framework in the Real Data

As we have already seen on the synthetic examples, Sect. 5.6 using linear regression / Granger causality has a comparably high, or higher power of the test (and ROC ratio) for data with linear causal structure, but it can perform no better than a random classifier when nonlinear causality is present. Below, we introduce a new set of experiments to analyse what happens if linear regression is applied to the real data. We build on the results for the commodity futures data, but for the purpose of clarity and compactness focus our attention on the causal relationships between the 1 month future contracts (1m WTI), and Baltic Dry Index (BDI). To ensure comparability, we are use the same setting

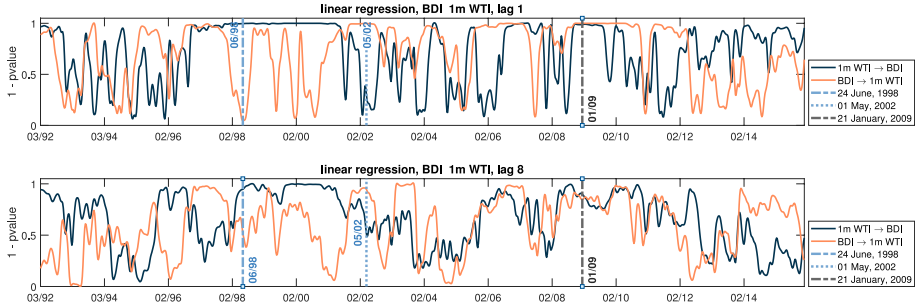

Fig. 28 Evolution of the causal influence tested with the linear regression (GCCA toolbox): 1-pvalues of the test statistic for 1 months WTI and BDI, with 1 lags (top subplot) and 8 lags (bottom subplot), rolling window of 104 weeks and cubic spline smoothing

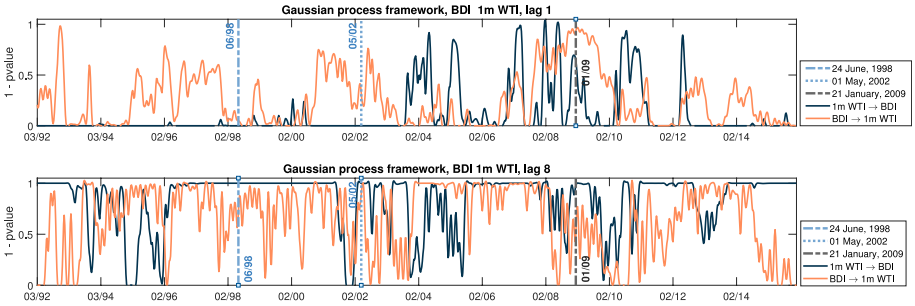

Fig. 29 Evolution of the causal influence tested with the framework based on GPs: 1-pvalues of the test statistic for 1 months WTI and BDI, with 1 lags (top subplot) and 8 lags (bottom subplot), rolling window of 104 weeks and cubic spline smoothing

as introduced in the Sect. 6.1: weekly data, lags 1 and 8, window length of 104, Matern covariance with 3 degrees of freedom, and the code we use for Granger causality is based on the GCCA toolbox Geweke (1982).

The results of the causality with linear regression are presented in the Fig. 28, which we contrast with the results for our framework in Fig. 29. These two approaches paint considerably different pictures for the causal relationships between the two time series. Fundamentally, at lag 1, the linear regression framework shows high confidence for the causality (precisely, for rejections of the hypothesis of lack of causality), which contrasts with GP framework rejecting lack of causality for very few data windows. We conjecture, that linear regression model is overconfident due to not being able to recognise nonlinear effects, in particular to remove excess serial correlation that would subsequently invalidate the assumptions of the hypothesis test resulting in excess kurtosis in the test statistic distribution and overly confident decision outcomes as a result. This is confirmed when analysing residuals of the linear regression fits. We demonstrate that for three specific point in time to show three scenarios where either one, or both of the directions show a high confidence for the linear model, that we observe with our framework.

Figure 30 presents a series Quantile-Quantile (QQ) plots of empirical residual quantiles versus normal quantiles of the residuals for the linear regression models for testing causality, and relate to the three dates marked on the evolution of causal influence in Fig. 29. Linear regression for the window ending on 24th January 1998 (first row in Fig. 30), strongly suggests a causal direction from 1 month futures contract to the Baltic Dry Index for 1 lag, a relationship which our framework strongly rejects. But when we look at the residuals of the linear model, we see evidence of serial correlation and skewness, and this is arguably stronger than for the opposite direction for which linear regression model does not support the existence of causality. For window ending on 1st May 2002, linear regression results with residuals that exhibit very strong leptokurtic tails in both directions – and again our framework does not support the hypothesis of lack of causality here. Finally for the window ending on 21st January 2009 linear regression again does not sufficiently account for serial correlation, but in this case our framework rejects the hypothesis of lack of causality for the direction of BDI to 1 month WTI.

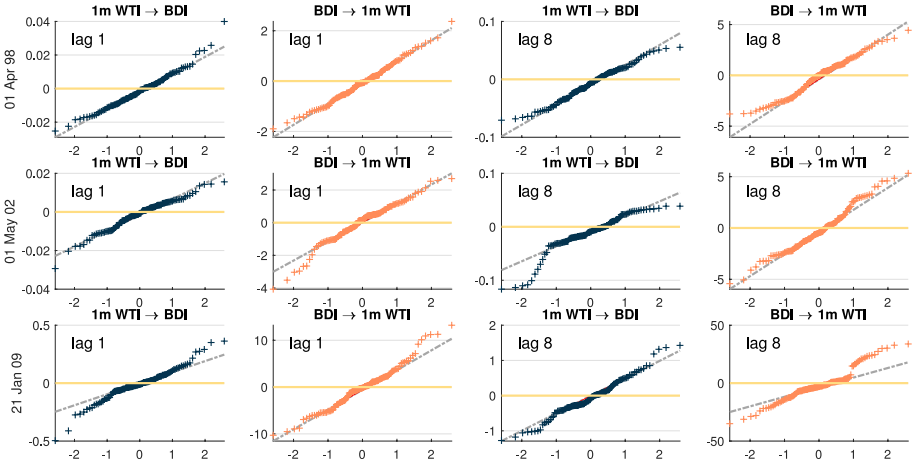

Fig. 30 QQ plots of the residuals for the linear regression models for testing causality, for data windows ending on: data windows ending on 24th January 1998, 1st May 2002, and 21st January 2009 (rows). Each of the four columns of qq plots represent a combination of lag and direction of the causality

Our conjecture of serial correlation in residuals leading to overconfidence of the linear model is supported by results that correct for such serial correlation. Figure 31 presents the result of testing for causality with a GP framework that a) incorporates linear trend from linear regression, and b) does not incorporate causal structure in the covariance, while the GP framework from Fig. 32 incorporates a) linear trend from the linear regression, b) allows for causality in covariance. Correcting for serial correlation removes some of the overconfidence of the linear regression model, which is then further reduced by also correcting for potential dependence in the covariance.

We conclude that while using linear regression models for testing causality can have higher power, this could be misleading, as the model could be overconfident due to incorrect statistical assumptions. Using GPs can not only help with these specific structural properties that we mentioned: serial correlation and causality in covariance, but it goes even further, by allowing to test for causality under a range of model assumptions without penalising model misspecification.

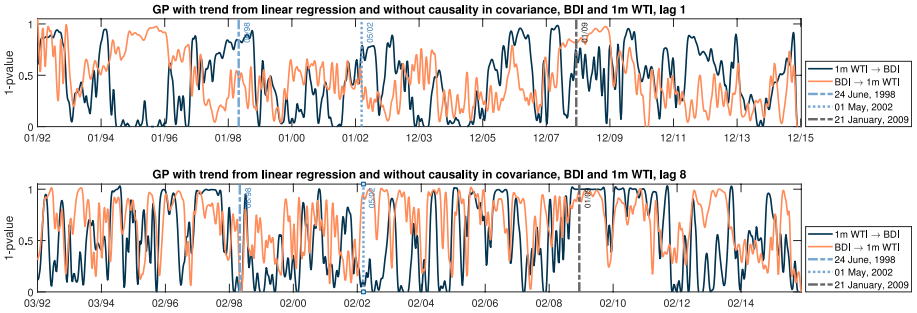

Fig. 31 Evolution of the causal influence tested with the framework based on GPs with trend from linear regression and no causality in covariance: 1-pvalues of the test statistic for 1 months WTI and BDI, with 1 lags (top subplot) and 8 lags (bottom subplot), rolling window of 104 weeks and cubic spline smoothing

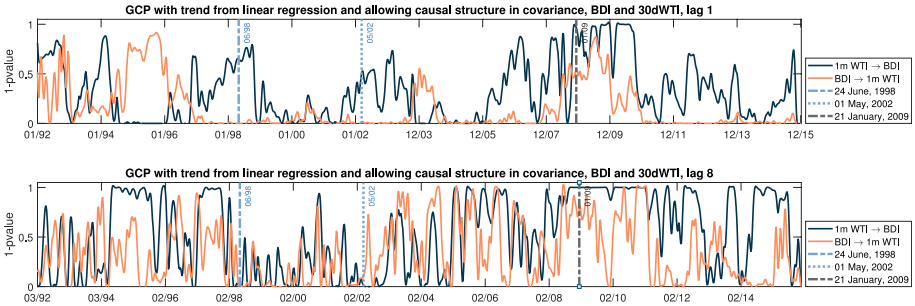

Fig. 32 Evolution of the causal influence tested with the framework based on GPs with trend from linear regression and allowing for causality in covariance: 1-pvalues of the test statistic for 1 months WTI and BDI, with 1 lags (top subplot) and 8 lags (bottom subplot), rolling window of 104 weeks and cubic spline smoothing

6.4 Real Data Experiment Conclusions

We summarise the results of the real data experiment, by revisiting our questions and remarks from the Sect. 6.1. Firstly, we conclude that 8 weeks is generally enough for each of the markets to price in associated causal impacts in both oil futures markets and currency markets, which supports the literature that relates to efficient market hypothesis. We conclude that the different classes of investments affect the type and speed of reaction. We also observe, that the direction and significance of causal influence is affected by regimes, as shown on the example of the period of low prices: 17.01.1990 – 11.08.1999 (bellow $40), and period of high prices: 26.05.2004 – 11.03.2009 (above $90).

Our analysis involved only three investment classes, and therefore is in no way sufficient to understand all important risk factors. We do however point out, that useful information can be obtained from analysing similarity of causal effects of two different factors. Such similarity can suggest that both factors are affected by a common factor (market volatility in our case). Increasing similarity of causal dependence can be understood in terms of systemic risk, see Billio et al. (2012).

7 Conclusion

We demonstrated that our proposed testing frameworks for statistical causality in general classes of multivariate nonlinear time series models are statistically efficient in detecting a wide range of different causality structures in complex multivariate nonlinear time series structures. It accommodates flexible features where causality may be present in either: trend, volatility or both structural components of the multivariate time series considered.

The analysis of the power of the hypothesis tests shows that the framework not only behaves as expected but also has properties that make it practical. An important result in this paper is obtaining a test statistic with known asymptotic distribution, but what is even more important is that we do not need a very large sample to be able to use that result in practice. For simple tests – ones that use exact hyperparameters, and compound tests – where the hyperparameters are estimated, we look at popular tools for assessing the quality of a testing procedure: test statistic distribution, power of the test and the ROC

curves. Furthermore, we compare our approach to Granger causality and transfer entropy – typical benchmarks for testing causality, and we conclude that our approach is practical in all cases, but offers superior performance especially for time series with long memory. Finally, we offer an example of real data application to analysing risk factors that investors should consider when building a portfolio of oil futures, currencies and physicals.

Additional Material

Calculating Marginal Likelihood

In the Eq. 14 we defined the causality metrics $L_{X\to Y|Z}$ as a logarithm of a likelihood ratio we have:

$$L_{X\to Y|Z} = \max_{\boldsymbol{\theta}_B} \ln p(\mathbf{Y} \mid \mathbf{X}^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m}; \boldsymbol{\theta}_B, \mathcal{M}_B) - \max_{\boldsymbol{\theta}_A} \ln p(\mathbf{Y} \mid \mathbf{Y}^{-l}, \mathbf{Z}^{-m}; \boldsymbol{\theta}_A, \mathcal{M}_A)$$

The two log-likelihood terms can be obtained in an analogous way, so lets show it for the model B. The Eq. 10 was defining $\mathbf{Y}_t$ as $\mathbf{Y}_t = f_B(\mathbf{X}_{t-1}^{-k}, \mathbf{Y}_{t-1}^{-l}, \mathbf{Z}_{t-1}^{-m}) + \epsilon_t^B$ with $\epsilon_t^B \sim \mathcal{N}(0, \sigma_B^2)$. If we denote $\mathbf{Y}$ as vector of all Y_t's, use $\boldsymbol{\mu}_B$ and $\mathbf{K}_B$ and $\boldsymbol{\Sigma}^B$ then we get the following distributions:

$$p(\mathbf{Y} \mid \mathbf{X}^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m}, f_B(\cdot)) = \mathcal{N}(\mathbf{Y}; f_B(\mathbf{X}^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m}), \boldsymbol{\Sigma}^B)$$
$$p(f_B(\cdot) \mid \mathbf{X}^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m}) = \mathcal{N}(f_B(\mathbf{X}^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m}); \boldsymbol{\mu}_B, \mathbf{K}_B)$$

Which are combined to calculate the marginal likelihood:

$$p(\mathbf{Y} \mid \mathbf{X}^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m}) = \int p(\mathbf{Y} \mid \mathbf{X}^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m}, f_B(\cdot)) p(f_B(\cdot) \mid \mathbf{X}^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m}) df_B(\cdot)$$

and that gives (for example by brute force and completing the squares) the following marginal likelihood:

$$p(\mathbf{Y} \mid \mathbf{X}^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m}; \boldsymbol{\theta}_B, \mathcal{M}_B) = \mathcal{N}(\mathbf{Y}; \boldsymbol{\mu}_B, \mathbf{K}_B + \boldsymbol{\Sigma}^B)$$

The marginal log-likelihood (or log marginal likelihood) is therefore equal:

$$\ln p(\mathbf{Y} \mid \mathbf{X}^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m}) = -\frac{1}{2}(\mathbf{Y} - \boldsymbol{\mu}_B)^{\mathrm{T}}(\mathbf{K}_B + \boldsymbol{\Sigma}^B)^{-1}(\mathbf{Y} - \boldsymbol{\mu}_B) - \frac{1}{2}\ln \mid \mathbf{K}_B + \boldsymbol{\Sigma}^B \mid -\frac{N}{2}\ln 2\pi.$$

What we need for the causality metrics in the Eq. 14 is the maximum marginal log-likelihood, which we can neatly achieve from Rasmussen and Williams (2006) as:

$$\frac{\partial}{\partial \boldsymbol{\theta}_j^B} \ln p(\mathbf{Y} \mid \mathbf{X}^{-k}, \mathbf{Y}^{-l}, \mathbf{Z}^{-m}; \boldsymbol{\theta}_B,) =$$
$$\frac{1}{2}(\mathbf{Y} - \boldsymbol{\mu}_B)^{\mathrm{T}}(\mathbf{K}_B + \boldsymbol{\Sigma}^B)^{-1}\frac{\partial(\mathbf{K}_B + \boldsymbol{\Sigma}^B)}{\partial \boldsymbol{\theta}_j^B}(\mathbf{K}_B + \boldsymbol{\Sigma}^B)^{-1}(\mathbf{Y} - \boldsymbol{\mu}_B) - \frac{1}{2}tr((\mathbf{K}_B + \boldsymbol{\Sigma}^B)^{-1}\frac{\partial(\mathbf{K}_B + \boldsymbol{\Sigma}^B)}{\partial \boldsymbol{\theta}_j^B})$$
$$= \frac{1}{2}\left(\left(\boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathrm{T}} - (\mathbf{K}_B + \boldsymbol{\Sigma}^B)^{-1}\right)\frac{\partial(\mathbf{K}_B + \boldsymbol{\Sigma}^B)}{\partial \boldsymbol{\theta}_j^B}\right), \text{ where } \boldsymbol{\alpha} = (\mathbf{K}_B + \boldsymbol{\Sigma}^B)^{-1}(\mathbf{Y} - \boldsymbol{\mu}_B).$$

Receiver Operating Characteristic (ROC)

Receiver operating characteristic (ROC) curves are commonly used in classification models to quantify the accuracy with which a model can discriminate between two classes. If we called one class as containing positive cases and the other – negative cases, then if the model correctly classifies, it will produce "true positive" and "true negative" labels, and if it incorrectly classifies, it will produce "false positive" and "false negative" cases.

The ROC curve plots True Positive Rate (TPR or Sensitivity) versus False Positive Rate (FPR or 1- Specificity), for a range of thresholds T:

$$TPR(T) = \frac{\sum \text{true positive}}{\sum \text{model positive (T)}} \quad FPR(T) = \frac{\sum \text{false positive}}{\sum \text{model negative (T)}}$$

In the case of continuous variables, as we have been dealing with, the classification rule is based on the test statistics being above / below a threshold, or it is cumulative distribution being above / below appropriate threshold. Recall the Sect. 4.1.2 and the asymptotic result for the distribution of the test statistics $2\hat{L}_{X\to Y|Z} \sim \chi^2_{2k}$. We were mentioning that the intuition is that large values of the causality "metrics" $L_{X\to Y|Z}$ coincide with causality, while lower – with lack of causality. So our classification rule could be to reject the hypothesis H_0 of no causality if $L_{X\to Y|Z} > 0.5T \Leftrightarrow 2L_{X\to Y|Z} > T$ and accordingly $F_{\chi^2_q}(2L_{X\to Y|Z}) > 1-\alpha$ for some related significance value of α.

Then let f_0 be a cdf of χ^2_q, while f_1 be a cdf of non-central χ^2_q. Then:

$$TPR(T) = \int_{-\infty}^{T} f_0(x)dx \quad FPR(T) = \int_{-\infty}^{T} f_1(x)dx.$$

The coordinates of the ROC curve are $(FPR(T), TPR(T))$, which leads to parametrisation:

$$\begin{cases} x = F_1(T) \\ y = F_0(F_1^{-1}(T)) \end{cases}$$

See Zou et al. (2007); Hillis and Metz (2012).

Data Availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

References

Alvarez MA, Lawrence ND (2011) Computationally efficient convolved multiple output gaussian processes. J Mach Learn Res 12:1459–1500

Amblard PO, Michel OJJ, Richard C, Honeine P (2012a) A Gaussian process regression approach for testing Granger causality between time series data. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3357–3360
Amblard PO, Vincent R, Michel OJJ, Richard C (2012b) Kernelizing Geweke's measures of granger causality. In: 2012 IEEE International Workshop on Machine Learning for Signal Processing. IEEE, pp 1–6
Ames M, Bagnarosa G, Peters GW, Shevchenko PV, Matsui T (2016) Which Risk Factors Drive Oil Futures Price Curves? Speculation and Hedging in the Short and Long-Term. Tech. Rep. ID 2840730, Social Science Research Network, Rochester, NY
Bakshi G, Panayotov G, Skoulakis G (2010) The Baltic Dry Index as a Predictor of Global Stock Returns, Commodity Returns, and Global Economic Activity. Tech. rep, Social Science Research Network, Rochester, NY
Barnett L, Bossomaier T (2012) Transfer Entropy as a Log-Likelihood Ratio. Phys Rev Lett 109(13)
Billio M, Getmansky M, Lo AW, Pelizzon L (2012) Econometric measures of connectedness and systemic risk in the finance and insurance sectors. J Financ Econ 104(3):535–559
Boyle P, Frean M (2005) Dependent Gaussian processes. In: In Advances in Neural Information Processing Systems 17, pp. 217–224. MIT Press
Campbell JY, Lo AW, MacKinlay AC (1997) The Econometrics of Financial Markets. Princeton University Press, Princeton
Campbell JY, Shiller RJ (1988) Stock Prices, Earnings, and Expected Dividends. J Financ 43(3):661–676
Chen WD (2006) Estimating the long memory granger causality effect with a spectrum estimator. J Forecast 25(3):193–200
Clarke KA (2000) Testing nonnested models of international relations: Reevaluating realism. Am J Pol Sci 45:724–744
Cressie N (1993) Statistics for Spatial Data, 2, edition. Wiley-Interscience, New York
Dempster MAH, Medova E, Tang K (2012) Determinants of oil futures prices and convenience yields. Quantitative Finance 12(12):1795–1809
Eichler M (2001) Granger causality graphs for multivariate time series
Eichler M (2001) Graphical modelling of multivariate time series. Tech Rep
Eichler M, Didelez V (2007) Causal reasoning in graphical time series models. In: Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence, pp. 109–116. AUAI Press
Fama EF (1970) Efficient Capital Markets: A Review of Theory and Empirical Work. J Finan 25(2):383–417
Fama EF, French K (1988) Permanent and Temporary Components of Stock Price. J Polit Econ 96(2):246–73
Fatichi S (2009) ARFIMA simulations. https://uk.mathworks.com/matlabcentral/fileexchange/25611-arfima-simulations
Garthwaite PH, Jolliffe I, Jones B (2002) Statistical Inference. Oxford; New York, OUP Oxford
Geweke J (1982) Measurement of Linear Dependence and Feedback between Multiple Time Series. J Am Stat Assoc 77(378):304–313
Granger CWJ (1963) Economic processes involving feedback. Inf Control 6(1):28–48
Granger CWJ (1980) Testing for causality : A personal viewpoint. J Econ Dyn Control 2(1):329–352
Hein M, Bousquet O (2004) Kernels, Associated Structures and Generalizations
Hillis SL, Metz CE (2012) An Analytic Expression for the Binormal Partial Area under the ROC Curve. Acad Radiol 19(12):1491–1498
Lungarella M, Ishiguro K, Kuniyoshi Y, Otsu N (2007) Methods for Quantifying the Causal Structure of Bivariate Time Series. Inte J Bifurcation Chaos 17(03):903–921
MacKay DJC (1994) Bayesian Non-linear Modelling for the Prediction Competition. In: In ASHRAE Transactions, V.100, Pt.2, pp. 1053–1062. ASHRAE
MacKinnon JG (1983) Model specification tests against non-nested alternatives. Economet Rev 2(1):85–110
Malkiel BG (2003) The Efficient Market Hypothesis and Its Critics. J Econ Pers 17:59–82
Melkumyan A, Ramos F (2009) A Sparse Covariance Function for Exact Gaussian Process Inference in Large Datasets. In: Proceedings of the 21st International Jont Conference on Artifical Intelligence, IJCAI'09, pp. 1936–1942. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA
Neal RM (1996) Bayesian Learning for Neural Networks, Lecture Notes in Statistics, vol. 118. Springer, New York, New York, NY
Pearl J (2000) Causality: models, reasoning, and inference. Cambridge University Press, New York, NY, USA
Pearl J (2010) An introduction to causal inference. Int J Biostat 6(2)
Pesaran MH, Weeks M (2001) Non-nested hypothesis testing: an overview. A companion to theoretical econometrics pp. 279–309

Qi Y, Minka TP, Picard RW, Ghahramani Z (2004) Predictive Automatic Relevance Determination by Expectation Propagation. In: Proceedings of the Twenty-first International Conference on Machine Learning, ICML '04, pp. 85–. ACM, New York, NY, USA

Rasmussen CE, Williams CKI (2006) Gaussian Processes for Machine Learning. MIT Press, Cambridge, Mass

Schoelkopf B, Tsuda K, Vert JP (2004) Kernel Methods in Computational Biology

Scholkopf B, Smola AJ (2001) Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA

Schreiber T (2000) Measuring Information Transfer. Phys Rev Lett 85(2):461–464

Snelson E, Ghahramani Z (2007) Local and global sparse Gaussian process approximations. AISTATS

Vuong QH (1089) Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica: J Econ Soc pp. 307–333

White H, Chalak K, Lu X (2011) Linking granger causality and the pearl causal model with settable systems. In: NIPS Mini-Symposium on Causality in Time Series, pp. 1–29

Wiener N (1956) The theory of prediction. In: Beckenbach Edwin F (ed) Modern Mathematics for Engineers, vol 1. McGraw-Hill, New York

Wilson P (2015) The misuse of the Vuong test for non-nested models to test for zero-inflation. Econ Lett 127:51–53

Zaremba A, Aste T (2014) Measures of Causality in Complex Datasets with Application to Financial Data. Entropy 16(4):2309–2349

Zou KH, O'Malley AJ, Mauri L (2007) Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. Circulation 115(5):654–657