Measuring Strength of Joint Causal Effects

Kurt Butler, *Graduate Student Member, IEEE,* Guanchao Feng, and Petar M. Djurić, *Life Fellow, IEEE*

***Abstract*—In the study of causality, we often seek not only to detect the presence of cause-effect relationships, but also to characterize how multiple causes combine to produce an effect. When the response to a change in one of the causes depends on the state of another cause, we say that there is an interaction or joint causation between the multiple causes. In this paper, we formalize a theory of joint causation based on higher-order derivatives and causal strength. Our proposed measure of joint causal strength is called the mixed differential causal effect (MDCE). We show that the MDCE approach can be naturally integrated into existing causal inference frameworks based on directed acyclic graphs or potential outcomes. We then derive a non-parametric estimator of the MDCE using Gaussian processes. We validate our approach with several experiments using synthetic data sets, demonstrating its applicability to static data as well as time series.**

***Index Terms*—causal effect, interaction, joint causality, Gaussian processes, nonlinear systems**

I. Introduction

Many scientific experiments study problems of causality, where one or more causal variables or factors combine to produce an effect. To describe cause-effect relationships rigorously, we employ statistical models [1]. Linear models are often a first choice for modeling, but linear models alone cannot describe *interactions* within a causal model, i.e., they fail to accurately describe situations in which the mechanism that one causal variable uses to produce an effect is moderated by the value of another causal variable [2]. In this case, the causal mechanism that produces the effect can only be understood when causes are considered jointly, and one cannot decompose the causal mechanism into independent pieces that separate the contributions of each causal variable.

Causality, as a rigorous statistical subject, has seen numerous different interpretations by different authors [3]–[5]. Interactions in causal models, as its own topic, has also received continuous interest throughout the past century [6]–[8]. This interest is also clearly seen in the applied sciences, including neuroscience [2], environmental science [9], psychology [10], economics [11], and epidemiology [12]. However the word *interaction* has a plurality of possible meanings [13], of which some do not align with our discussion here. For this reason, we prefer the term *joint causation* to express the idea that multiple causes produce an effect jointly, in an irreducible manner.

Most popular approaches to modeling interactions use parametric models, including bilinear models [14], generalized linear models [15], ensembles [16], or Volterra series [17], [18]. While parametric models are easy to analyze, each parametric family describes a different notion of interaction, which might differ from the notion produced by using another family of models. Additionally, the accurate estimation of joint causation depends on the expressiveness of the model class and preprocessing techniques, such as centering [2], [19]. A number of methods based on analysis of variance (ANOVA) have also been applied to discover interactions [20]. The ANOVA approach has a particular usefulness in that it can describe categorical causal variables. However, ANOVA also can be formulated as a linear regression model [21]. A more properly non-parametric approach is the Sobol' method [22], which considers an orthogonal decomposition of the function of interest into pieces. each utilizing a subset of the input variables. The amount of variances contributed by a set of variables can then be used as a measure of joint sensitivity, called Sobol' indices [23]. These indices are powerful as they can quantify interactions in a model-free manner. However, being a global method, this approach cannot provide local information about the model.

The contribution of this paper is to propose a principled and non-parametric theory of joint causation using higher-order derivatives and causal strength. Our proposed measure, the mixed differential causal effect (MDCE), represents the joint causation as a function that expresses how multiple causal variables interact to produce a given effect. The MDCE can be effectively estimated using Gaussian processes (GPs) to produce a Bayesian posterior over functions. Several advantages of the GP approach are examined through examples, including the detection of joint causation, robustness to change of kernel, sparse approximation of the GP kernel, and applications to time series.

We organize the paper as follows. We present literature related to this work in Section II. In Section III, we introduce the problem of joint causation and the MDCE approach to studying it. In Section IV, we estimate the MDCE from data using Gaussian process regression (GPR), and we discuss several aspects of this approach theory, including robustness to the kernel choice, detection of joint causation and approximations to exact GPR. We demonstrate the performance of our approach in Section V, and we provide conclusions in Section VI.

II. Related Work

Potentially the most popular family of interaction-based models are Volterra models [17], [18], [24]. Volterra models are restrictive in the sense that specification of the Volterra kernel also specifies the notion of interaction under study, and changing the Volterra kernel changes the notion of interaction.

This work was supported by the National Science Foundation under Award 2212506. The work of Kurt Butler was supported by the U.S. Department of Education under the GAANN Fellowship.

The authors are with the Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY 11794-2350 USA (e-mail: kurt.butler@stonybrook.edu; guanchao.feng@stonybrook.edu; petar.djuric@stonybrook.edu).

An advantage of our differential approach is that it is straightforward to compare the MDCE estimated via GPR to what we would expect under a particular Volterra model. Furthermore, the notion of MDCE also provides a natural way to compare the results obtained using distinct Volterra kernels.

Designing a GP kernel to account for interactions has also been proposed before [25]–[27]. These approaches invoke an additive decomposition of the GP kernels to produce a GP predictor that is a sum of interacting and non-interacting parts. Support vector ANOVA [28] provides a separate but related framework to produce a similar decomposition. Like Volterra modeling, the kernel decomposition also specifies the notion of interaction, but the interpretation is less straightforward because distinct kernels can approximate the same function. However, we show that additive kernels can be combined with the MDCE approach in Section V.

III. Problem Formulation

In this section, we introduce our theory of causal strength and joint causation and discuss several interesting aspects of it. We begin by framing our work in the context of the existing causality literature.

A. Paradigms of causal inference

There are various approaches to modern causal inference, including the graphical framework of Pearl [5] and the potential outcomes framework of Rubin [4]. Although they approach problems differently, the frameworks of Pearl and Rubin are often interpreted to represent a common defintion of causality, called *interventional causality* [5, p.243-245]. We distinguish this from a third popular framework, Granger causality [3], which is well-established in econometrics and neuroscience [29]. However, it should be noted that Granger causality and interventional causality are not equivalent ideas.[1]

Regardless of one's framework and definition of causality, our proposed theory of joint causation can be employed to explain how multiple causes produce their effects. Our theory is best understood using the concept of a **causal mechanism** [31], that is, a function F that takes in a vector of cause variables $\mathbf{x} = (x_1, ..., x_D)$ and a noise variable ε and assigns a value to a response variable y, denoted as $y := F(\mathbf{x}, \varepsilon)$. Causal strength and joint causation are then properties of a causal mechanism. To relate our approach to the existing schools of thought, we propose defining a causal mechanism within the frameworks of Pearl, Rubin, and Granger.

1) Causal graphs: In the Pearl framework, one models the causal relationships between a set of random variables $X_1, ..., X_D$ using a directed acyclic graph (DAG) $\mathcal{G}$ which encodes the conditional dependencies between variables [5]. A structural causal model (SCM) extends the DAG model by providing an explicit description of how each X_i obtains its value [30]. Namely, each X_i is determined by a functional assignment

$$X_i := F_i(\mathbf{Pa}_i, \varepsilon_i),$$

where $\mathbf{Pa}_i = [X_{j_1}, ..., X_{j_{K_i}}]$ are the parents of X_i in $\mathcal{G}$, $K_i = \dim(\mathbf{Pa}_i)$, and ε_i is an exogenous noise term.

In this setting, each F_i is a causal mechanism, where $\mathbf{Pa}_i$ are the input causes and X_i is the response variable. In our analysis, we want to measure how strongly each parent $X_j \in \mathbf{Pa}_i$ influences X_i, as viewed through the particular mechanism F_i. This approach measures the causal strength of a parent-child relationship in the DAG, which we refer to as *direct causal effect*. We distinguish this from *total causal effect*, where the effect of X_j on X_i is calculated by summing over every possible path in the DAG [32]. Total causal effect thus considers that many causal mechanisms may be composed to go from cause to effect. In this paper, we only work with direct causal effect, since questions about the total causal effect may be derived from the direct causal effect [33].

2) Potential outcomes: In the Rubin framework, we work with a vector of D possible causes $\mathbf{a} = [a_1, ..., a_D]$, and a potential outcome function $y_n(\mathbf{a})$ for each instance of the response variable [34]. Our goal in this framework is to characterize the distribution of $Y_n(\mathbf{a})$ for every possible $\mathbf{a}$.

To express the potential outcome function as a causal mechanism, we let $\mathbf{x} = \mathbf{a}$, and we use the reparameterization trick [35] to express the random variable $Y_n(\mathbf{x})$ as a function $F(\mathbf{x}, \varepsilon_n)$ of a noise variable ε_n.

3) Granger causality: Granger causality relies on two principles: 1. the cause precedes its effect, and 2. the cause possesses unique information about the future values of its effect [36]. If a time series, denoted as x, is considered to cause another time series, denoted as y, then forecasts of y that incorporate both its own past values and the past values of x demonstrate greater accuracy compared to predictions based solely on the past values of y. In our context, if y_{t+1} represents a sample of the time series y observed at time instant $t+1$, using a functional assignment, our interest is in

$$y_{t+1} := F(\mathbf{Pa}_t, \varepsilon_{t+1}), \tag{1}$$

where $\mathbf{Pa}_t$ represents the parents of y_{t+1}, which could be past samples of the time series y and x.

B. Causal strength and joint causation

To begin, consider a causal mechanism that assigns a value to the response variable y,

$$y := F(x_1, ..., x_D) + \varepsilon,$$

where $\mathbf{x} = [x_1, ..., x_D]$ is a set of observed cause variables and ε is a noise variable, which we now assume to be additive.[2] The symbol $:=$ is used to indicate the direction of causality [30]. Typical examples of this model include the *linear model*,

$$y := \sum_{i=1}^{D} a_i x_i + \varepsilon, \tag{2}$$

[1]Granger causality is defined using predictive ability and the so-called arrow of time. This differs fundamentally from interventional causality, which defines causality via *interventions* upon a system. It is possible that Granger causal analysis can be useful when working in an interventional framework [30, p.201], but without limiting assumptions, neither Granger causality nor interventional causality can logically imply the other.

[2]Additive noise is not required for defining causal strength, but all models we consider will use this assumption.

and the *bilinear* or quadratic model,

$$y := \sum_{i=1}^{D} a_i x_i + \sum_{i,j=1}^{D} b_{ij} x_i x_j + \varepsilon, \tag{3}$$

where a_i, b_{ij} are constants in each model.

When the function F is continuously differentiable, the partial derivatives $\partial F/\partial x_i$ quantify the sensitivity of the effect variable to local changes in each cause x_i. In the context of causal models, this can also be interpreted as a measure of the strength of causation [37]. For this reason, we define the **differential causal effect** (DCE) of x_i on y to be the partial derivative,

$$\text{DCE}_{x_i \to y}(x_1, ..., x_D) = \frac{\partial F(x_1, ..., x_D)}{\partial x_i}. \tag{4}$$

We observe in (4) that it is possible for $\text{DCE}_{x_i \to y}$ to be a function of x_j ($i \neq j$), meaning that the causal strength of x_i on y is being moderated by another variable x_j. We introduce the shorthand notation

$$\partial_i \triangleq \frac{\partial}{\partial x_i}$$

to make some of the expressions more concise.

Inversely, one notion of independence of multiple causes is that the DCE of each x_i on y only depends upon x_i; i.e., $\partial_i F$ is only a function of x_i. As a result, the mixed second-derivatives of F are zero,

$$\frac{\partial^2 F}{\partial x_j \partial x_i} = \frac{\partial}{\partial x_j}\left(\frac{\partial F(x_1, ..., x_D)}{\partial x_i}\right) = 0, \qquad i \neq j. \tag{5}$$

The quantity $\partial_j \partial_i F$ in (5), which we will call the **mixed differential causal effect** (MDCE), thus describes the manner in which x_j moderates the strength of causation $x_i \to y$. The sign of the MDCE can describe whether the interaction of x_i and x_j is 'synergistic' or 'antagonistic' [8]. We may also denote the MDCE by $\partial^2 y/\partial x_i \partial x_j$ or $\partial_i \partial_j y$ when the causal mechanism F is implicitly understood. Note that since

$$\partial_i \partial_j F = \partial_j \partial_i F \tag{6}$$

whenever F has continuous second derivatives [38], the joint causation is usually symmetric, i.e., the ordering of i, j does not matter.

We say that x_i, x_j are *separable* causes of y if the function F admits an additive decomposition that separates x_i and x_j:

$$F(x_i, x_j, \mathbf{x}_{-i,-j}) = F_1(x_i, \mathbf{x}_{-i}) + F_2(x_j, \mathbf{x}_{-j}) \tag{7}$$

where the vector $\mathbf{x}_{-i,-j}$ contains all the variables x_k except x_i and x_j (similarly, $\mathbf{x}_{-i}$ and $\mathbf{x}_{-j}$ represent the vectors excluding x_i and x_j, respectively). When F has continuous second derivatives, x_i and x_j are separable if and only if $\partial_i \partial_j F = 0$. Otherwise, we say that the pair x_i, x_j *jointly causes* y if $\partial_j \partial_i F$ is not identically zero. When x_i and x_j jointly cause y, it is impossible to discuss the causal effect of x_i on y without also considering x_j, and vice-versa. For example in the bilinear model (3), the mixed derivatives are

$$\frac{\partial^2 F}{\partial x_i \partial x_j} = b_{ij} + b_{ji}.$$

We see immediately that the mixed derivatives for the bilinear model are constants, and unchanged if we switch i and j. The magnitude of $\partial_i \partial_j F$ quantifies the extent to which x_i, x_j are interacting when they drive y.

In order to study joint causation in practice, one must be able to estimate derivatives of functions from data, which we address in Section IV.

C. Higher-order joint causation

To discuss the joint causation of three or more causes, we extend the analysis in an obvious way. We define the higher-order mixed derivatives $\partial_{i_1} \partial_{i_2} \cdots \partial_{i_K} F$ to be the K-th order MDCE, which describes the joint causation of $x_{i_1}, x_{i_2}, ..., x_{i_K}$ on y. Since $\partial_{i_1} \cdots \partial_{i_K} F$ is only nonzero if $\partial_{i_a} \partial_{i_b} F \neq 0$, for all $i_a, i_b \in \{i_1, ..., i_K\}$ s.t. $i_a \neq i_b$, pairwise joint causation is a necessary condition for higher-order joint causation.

D. Change of variables

We have already noted some elementary properties of the MDCE, such as symmetry (6) and the additive decomposition formula (7). Some additional important properties arise when considering a change of variables.

Joint causation is not invariant to nonlinear transformations at the output; that is, we can expect the joint causation to differ if we change the output quantity of interest. As an elementary example, consider the following causal system:

$$\begin{aligned} x_1 &\sim \mathcal{U}(0,1) \\ x_2 &\sim \mathcal{U}(0,1) \\ y &:= a x_1 + b x_2 \\ \tilde{y} &= y^2, \end{aligned}$$

where $a, b > 0$. For the sake of causality, we regard y and $\tilde{y}$ as different representations of *the same variable* in the system, so that we are not discussing chains of causal interactions $x_i \to y \to \tilde{y}$. An analogy would be that standard deviation and variance are equivalent quantities, described in different coordinates.

The joint causation of x_1 and x_2 on y is zero, since the functional relationship between the three is linear. However, we observe that because

$$\tilde{y} = a^2 x_1^2 + b^2 x_2^2 + 2ab x_1 x_2,$$

the joint causation of x_1 and x_2 on $\tilde{y}$ is non-zero, since the MDCE is the constant function $2ab$. We remark that this example not only applies to the MDCE, but also to traditional measures of interaction as well.

To understand this phenomenon in general, consider the situation in which $y := F(x_1, x_2)$ is a causal mechanism, and we have a nonlinear transformation at the output, $\tilde{y} = g(y)$. Ordinary calculus tells us how to relate $\partial_1 \partial_2 y$ and $\partial_1 \partial_2 \tilde{y}$:

$$\underbrace{\frac{\partial^2 \tilde{y}}{\partial x_1 \partial x_2}}_{\text{Transformed MDCE}} = \underbrace{\frac{\partial^2 \tilde{y}}{\partial y^2}\frac{\partial y}{\partial x_1}\frac{\partial y}{\partial x_2}}_{\text{Bias term}} + \frac{\partial \tilde{y}}{\partial y}\underbrace{\frac{\partial^2 y}{\partial x_1 \partial x_2}}_{\text{Original MDCE}}. \tag{8}$$

The second term in (8) contains the original MDCE, scaled by a function $\partial \tilde{y}/\partial y$ analogous to the chain rule for the first

derivative. However, the bias term in (8) will distort the MDCE estimate when we transform y nonlinearly. As a result, joint causation is not invariant under reparameterization, and the previous example shows that while we may find no joint causation (or interaction) in one coordinate system, there may be a nonzero interaction in a different coordinate system.

For transformations of the cause variables, the situation is much more tame. If $x_1 = g_1(\tilde{x}_1)$ and $x_2 = g_2(\tilde{x}_2)$, for some appropriate functions g_1, g_2, then the MDCE is only scaled according to the coordinate change,

$$\frac{\partial^2 y}{\partial \tilde{x}_1 \partial \tilde{x}_2} = \frac{\partial^2 y}{\partial x_1 \partial x_2} \frac{\partial x_1}{\partial \tilde{x}_1} \frac{\partial x_2}{\partial \tilde{x}_2}. \tag{9}$$

As a result, if the joint causation of x_1, x_2 on y is zero, it remains zero for any reparameterization of these variables.

E. Confounders

When studying causal systems in the real world, we must be cautious about the presence of unobserved *confounders*, variables that exert a causal influence on multiple observed covariates [5, p.12]. In our discussion so far, we have focused on the analysis of individual causal mechanisms. However, confounders distort our inference because they affect multiple causal mechanisms simultaneously. A basic confounding scenario is the "common cause" situation, in which a vector of unobserved variables $\mathbf{z}$ exerts a causal influence on both $\mathbf{x}$ and y. To describe the situation in a model, we would need at least two causal mechanisms,

$$\mathbf{x} := G(\mathbf{z}) + \boldsymbol{\varepsilon}_x, \tag{10}$$

$$y := F(\mathbf{x}, \mathbf{z}) + \varepsilon_y. \tag{11}$$

When the confounder $\mathbf{z}$ influences both $\mathbf{x}$ and y, it becomes impossible to distinguish when a change in y is directly due to the change in $\mathbf{x}$, or if both changes were due the common cause $\mathbf{z}$. As a result, the presence of a confounder can sabotage our measurements of causal strength. In particular, the two issues are that (1) confounders could prevent us from accurately estimating the mechanism F , and (2) the DCE or MDCE could depend on the value of $\mathbf{z}$. Simpson's paradox demonstrates that due to confounders, the true causal strength could be dramatically different from what we measure [37].

Although confounders can create very challenging scenarios for measuring causal strengths, under certain assumptions it may be possible to estimate MDCEs. In Section V-C, we study a system where the MDCE can still be estimated efficiently despite the influence of a so-called 'linear confounder.' While this experiment is curious, we leave a more systematic study of confounder bias for future work.

IV. Discovering causation when F is unknown

A. Gaussian process regression

We now review Gaussian process regression (GPR) as a non-parametric, probabilistic tool to estimate functions, and their higher-order derivatives, from data. We are given a vector of covariates $\mathbf{x} = (x_1, x_2, ..., x_D)$ and a target quantity y, given by

$$y = F(\mathbf{x}) + \varepsilon,$$

where $F(\cdot)$ is unknown, and ε is zero-mean white Gaussian noise with variance σ^2, which we assume to be fixed for now. GPR estimates the function F in a Bayesian manner by placing a Gaussian process (GP) prior over the space of possible functions.

We say that a function F is a GP if for any finite set of points $\mathbf{x}_1, ..., \mathbf{x}_N$, the probability distribution over $F(\mathbf{x}_1), ..., F(\mathbf{x}_N)$ is multivariate normal. The function

$$m(\mathbf{x}) \triangleq \mathbb{E}(F(\mathbf{x}))$$

is called the *mean function*, and the function

$$k(\mathbf{x}, \mathbf{x}') \triangleq \mathrm{cov}(F(\mathbf{x}), F(\mathbf{x}'))$$

is referred to as *kernel* or *covariance function* of the GP. Together m and k uniquely specify a GP distribution. We write $F \sim \mathcal{GP}(m, k)$.

To learn functions from data, GPR takes a Bayesian approach. First, a GP prior is placed over F. When picking a prior, we often set $m(\mathbf{x}) \equiv 0$ in the absence of prior knowledge about $\mathbb{E}(F(\mathbf{x}))$. We then consider a data set $\{(\mathbf{x}_n, y_n); n = 1, ..., N\}$ consisting of input-output pairs. For notation, we let $\mathbf{X}$ be an $N \times D$ matrix whose nth row is $\mathbf{x}_n$, and $\mathbf{y}$ is an $N \times 1$ vector whose nth element is y_n. To make predictions about $F(\mathbf{x})$ at a new location $\mathbf{x}$, we obtain a posterior distribution after conditioning on $\mathbf{X}, \mathbf{y}, \mathbf{x}$, which is also Gaussian. Following the derivation in [25], we may express this posterior as

$$F(\mathbf{x})|\mathbf{X}, \mathbf{y}, \mathbf{x} \sim \mathcal{N}(m_p(\mathbf{x}), k_p(\mathbf{x}, \mathbf{x})), \tag{12}$$

where

$$m_p(\mathbf{x}) = \mathbf{k}_*(\mathbf{x})^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y}, \tag{13}$$

$$k_p(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - \mathbf{k}_*(\mathbf{x})^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{k}_*(\mathbf{x}'), \tag{14}$$

are the posterior mean and kernel, respectively, and σ^2 is the noise variance. The vector $\mathbf{k}_*(\mathbf{x})$ and the matrix $\mathbf{K}$ are commonly used notation in these formulas, and are given by

$$(\mathbf{k}_*(\mathbf{x}))_n = k(\mathbf{x}_n, \mathbf{x}), \qquad \mathbf{K}_{nn'} = k(\mathbf{x}_n, \mathbf{x}_{n'}).$$

The estimated function $\hat{F}$ is usually taken to be the posterior mean, m_p. Examining (13), we observe that $\hat{F}$ can be expressed in terms of a sum,

$$\hat{F}(\mathbf{x}) = m_p(\mathbf{x}) = \mathbf{k}_*(\mathbf{x})^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y} \tag{15}$$

$$= \sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) \alpha_n, \tag{16}$$

where $\boldsymbol{\alpha} = (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y}$ is a vector of constants.

For GPR to yield a meaningful predictive distribution, proper choices must be made in selecting the kernel function k and noise variance σ^2. Typically, k is selected from a parameterized family of kernels, where the kernel parameters and σ^2 are jointly optimized during training to maximize the likelihood of the resulting model [39]. We discuss kernel selection in detail in Section IV-C.

B. Differentiation of Gaussian processes

When the kernel function k of the GP prior has certain properties, such as continuity, differentiability or periodicity, then these properties will be conferred to samples from the GP posterior [40], [41]. In particular, if

$$F \sim \mathcal{GP}(m, k), \tag{17}$$

for appropriate functions m, k, one can show that $\partial_i F$ is also a Gaussian process [42]:

$$\frac{\partial F}{\partial x_i} \sim \mathcal{GP}\left(\frac{\partial m}{\partial x_i}, \frac{\partial^2 k}{\partial x_i \partial x_i'}\right). \tag{18}$$

From (18), we see that for F to be differentiable (as a function randomly sampled from the GP), it is a necessary condition that m and k are once and twice differentiable, respectively. This condition is also sufficient for the GP to be differentiable [41]. The equations (17) and (18) allow us to obtain a posterior over the derivatives, even if only observations of F are available. This result also shows that, when everything is defined, the mean and partial derivative commute: $\partial_i \mathbb{E}(F) = \mathbb{E}(\partial_i F)$.

To estimate derivatives of a function F from data, we can first estimate the posterior GP, $F|\mathbf{X}, \mathbf{y}$. Then given the GP, we can use (18) and (16) to get an estimator of the derivative, i.e.,

$$\widehat{\frac{\partial F(\mathbf{x})}{\partial x_i}} = \mathbb{E}\left(\frac{\partial F(\mathbf{x})}{\partial x_i}\middle|\mathbf{X}, \mathbf{y}, \mathbf{x}\right) \tag{19}$$

$$= \frac{\partial m_p}{\partial x_i} \tag{20}$$

$$= \sum_{n=1}^{N} \frac{\partial k(\mathbf{x}, \mathbf{x}_n)}{\partial x_i} \alpha_n. \tag{21}$$

Naturally, if we wish to study the MDCE using GPs, then we must obtain the posterior distribution of $\partial_i \partial_j F$. First, we must obtain the posterior distribution for the function F from data, as in (12), and then repeated application of (18) yields

$$\frac{\partial^2 F}{\partial x_i \partial x_j}\middle|\mathbf{X}, \mathbf{y} \sim \mathcal{GP}\left(\frac{\partial^2 m_p}{\partial x_i \partial x_j}, \frac{\partial^4 k_p}{\partial x_i \partial x_i' \partial x_j \partial x_j'}\right). \tag{22}$$

Due to the form of $\hat{F}$ given in (13), computing these quantities is straightforward. In particular,

$$\mathbb{E}\left(\frac{\partial^2 F}{\partial x_i \partial x_j}\middle|\mathbf{X}, \mathbf{y}\right) = \frac{\partial^2 m_p}{\partial x_i \partial x_j} = \sum_{n=1}^{N} \frac{\partial^2 k(\mathbf{x}, \mathbf{x}_n)}{\partial x_i \partial x_j} \alpha_n, \tag{23}$$

is an estimator of the mean of the MDCE. The quality of the estimator depends on the data $\mathbf{X}, \mathbf{y}$ and the choice of kernel k, which we now address.

C. Covariance functions

So far, we have assumed that the kernel for the GP prior, k, has been given. In practice, we adopt a parametric form for the kernel k, and we select the optimal parameters $\boldsymbol{\theta}$ by the maximum likelihood estimator [39]. The kernel parameters are often called *hyperparameters* because they do not directly determine the structure of the output function, but rather they tend to modify how the training data were used to produce a posterior. In general, the family of kernels being used is highly customizable, and designing a good kernel family can lead to a great deal of expressiveness in the GP model [25].

To produce the MDCE estimator as in (22), we will need to differentiate the kernel functions used by the GP. We will provide analytical results for some common kernels. For more complex kernels, we note that the use of computational tools, such as automatic differentiation [43] or symbolic differentiation, can greatly simplify the implementation of MDCE estimators in practice.

SE and ARD-SE kernels. The default kernel is often the squared-exponential (SE) kernel [25],

$$k_{\text{SE}}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\sum_{j=1}^{D} \frac{(x_j - x_j')^2}{2\ell^2}\right), \tag{24}$$

where σ_f, ℓ are hyperparameters and are called signal variance and length-scale, respectively. A useful and simple generalization over the SE kernel is the SE kernel with automatic relevance determination (ARD-SE) [44],

$$k_{\text{ARD-SE}}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\sum_{j=1}^{D} \frac{(x_j - x_j')^2}{2\ell_j^2}\right), \tag{25}$$

where each input dimension j receives its own length-scale parameter ℓ_j.

The SE and ARD-SE kernels can be differentiated easily. To make the calculation straightforward, notice that $k_{\text{ARD-SE}}$ and k_{SE} can be expressed as a product of 1-dimensional SE kernels. As a result, when $i \neq j$, it can be shown that

$$\frac{\partial^2 k_{\text{ARD-SE}}(\mathbf{x}, \mathbf{x}')}{\partial x_i \partial x_j} = \sigma_f^2 \exp\left(-\sum_{d=1}^{D} \frac{(x_d - x_d')^2}{2\ell_d^2}\right) \times \left(\frac{x_i - x_i'}{\ell_i^2}\right)\left(\frac{x_j - x_j'}{\ell_j^2}\right). \tag{26}$$

Using the same trick, we may compactly express the higher-order mixed derivatives as

$$\frac{\partial^L k_{\text{ARD-SE}}(\mathbf{x}, \mathbf{x}')}{\partial x_{i_1} \cdots \partial x_{i_L}} = k_{\text{ARD-SE}}(\mathbf{x}, \mathbf{x}') \prod_{l=1}^{L} \left(\frac{x_{i_l} - x_{i_l}'}{-\ell_{i_l}^2}\right),$$

where $i_1, ..., i_L$ is a list of indices with no repeats.

Periodic kernels. As noted earlier, other kernels are often of interest to enforce specific properties on the learned functions. The periodic kernel is sometimes used to enforce periodicity of the learned functions along each input dimension [45]:

$$k_{per}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{D} \frac{\sin^2(f_d(x_d - x_d'))}{\ell_d^2}\right), \tag{27}$$

where σ_f^2, r_i, f_i are again hyperparameters. The mixed derivative is again straightforward to compute, although the expressions get a bit more unwieldy,

$$\frac{\partial^2 k_{per}(\mathbf{x},\mathbf{x}')}{\partial x_i \partial x_j} = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{D}\frac{\sin^2(f_d(x_d - x'_d))}{\ell_d^2}\right) \times \left(\frac{f_i \sin(f_i(x_i - x'_i))\cos(f_i(x_i - x'_i))}{-\ell_i^2}\right) \times \left(\frac{f_j \sin(f_j(x_j - x'_j))\cos(f_j(x_j - x'_j))}{-\ell_j^2}\right).$$

Matérn kernels. The SE, ARD-SE and periodic kernels are smooth functions, meaning that they are infinitely differentiable, and as a result the GP posterior will model F as an infinitely-differentiable function. In cases where only functions of limited differentiability are of interest, it is common to consider the Matérn class of kernels [25], [46]. The general class of Matérn kernels is given by

$$k_{\text{Matern}}(r) = \sigma_f^2 \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}r}{\ell}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}r}{\ell}\right), \quad (28)$$

where $r \triangleq ||\mathbf{x} - \mathbf{x}'||$ is the Euclidean distance between $\mathbf{x}$ and $\mathbf{x}'$, and K_ν is a modified Bessel function. The hyperparameters ν, σ_f^2 and ℓ are positive numbers. When using a Matérn kernel, a function F sampled from the kernel is only k-times differentiable if and only if $k < \nu$. As a result, twice-differentiable functions can be modeled by selecting $\nu = 5/2$, leading to the Matérn 5/2 kernel,

$$k_{\text{Mat5/2}}(r) = \sigma_f^2\left(1 + \frac{\sqrt{5}r}{\ell} + \frac{5r^2}{3\ell^2}\right)\exp\left(-\frac{\sqrt{5}r}{\ell}\right), \quad (29)$$

where now ℓ and σ_f^2 are the only hyperparameters. For studying pairwise joint causation, at least two derivatives are required to compute the MDCE. The kernel (29) can be differentiated to yield

$$\frac{\partial^2 k_{\text{Mat5/2}}(\mathbf{x},\mathbf{x}')}{\partial x_i \partial x_j} = -\sigma_f^2\left(\frac{5\sqrt{5}r^2}{3\ell^3} + \frac{5r}{3\ell^2}\right)\exp\left(\frac{-\sqrt{5}r}{\ell}\right) \times \frac{(x_i - x'_i)(x_j - x'_j)}{r^2} + \sigma_f^2\left(\frac{25r^2}{3\ell^4} - \frac{5\sqrt{5}r}{3\ell^3} - \frac{5}{3\ell^2}\right) \times \exp\left(\frac{-\sqrt{5}r}{\ell}\right)\frac{(x_i - x'_i)(x_j - x'_j)}{-r^3},$$

which is somewhat less elegant but nonetheless tractable.

Combining old kernels to make new ones. The power of GPR comes not only from the wide number of kernels, but also from the ability to design new kernels through the summation and multiplication of existing kernels [40]. When kernels are added together, the resulting GP estimate also admits an additive decomposition. For example, if $k(x_1, x_2, x'_1, x'_2) = k_1(x_1, x'_1) + k_2(x_2, x'_2)$, then a function F sampled from the GP posterior will admit an additive decomposition as well; in this case, $F(\mathbf{x})$ can be expressed as $F_1(x_1) + F_2(x_2)$ [25].

The additive kernel in [26] combines several SE kernels together to produce such an additive decomposition of functions. The SE additive kernel for D input features may be summarized as

$$k_{\text{add}}(\mathbf{x},\mathbf{x}') = \sum_{i=1}^{D} k_{\text{SE}}(x_i, x'_i) + \sum_{i=1}^{D}\sum_{j=i+1}^{D} k_{\text{SE}}(x_i, x_j, x'_i, x'_j) \vdots + k_{\text{SE}}(\mathbf{x},\mathbf{x}'). \quad (30)$$

By linearity, any derivative of the additive kernel can be obtained by individually differentiating each of the kernels in the sum. Since we are using SE kernels in this model, the MDCE estimator given by this additive kernel can be directly obtained using (26).

D. Bayesian detection of joint causation

In this section, we consider the problem of deciding when x_i, x_j do not jointly cause y. To formulate the problem using decision theory, we consider a Bayesian multiple hypothesis testing framework [47]. Aside from decision making, knowledge of which joint causalities are detectable can be leveraged in kernel design to refine the GPR model.

Suppose that we have a set of candidate models $\mathcal{M}_0, \cdots, \mathcal{M}_{Q-1}$ to describe a data set $\mathbf{X}, \mathbf{y}$, and we assign prior probabilities $P(\mathcal{M}_0), \cdots, P(\mathcal{M}_{Q-1})$ to each model. Assuming that $\mathbf{X}$ and $\mathbf{y}$ were generated by one of the models, the minimum probability of error detector will select the model with the highest posterior probability [47]. Mathematically, we decide model $\mathcal{M}_r$ is the best model if

$$P(\mathcal{M}_r|\mathbf{X},\mathbf{y}) > P(\mathcal{M}_q|\mathbf{X},\mathbf{y}), \ \ \forall q \neq r. \quad (31)$$

To decide if a given pair, x_i, x_j jointly cause y, a binary test will suffice. In model $\mathcal{M}_0$, we suppose that the joint causation is null, $\partial_i\partial_j F \equiv 0$. According to (7), the function F admits a factorization that separates x_i and x_j. To encode this assumption into the model, we can enforce this constraint through the kernel function. Thus, under model $\mathcal{M}_0$, we have that

$$F|\mathcal{M}_0 \sim \mathcal{GP}(0, k), \quad (32)$$

$$k(\mathbf{x},\mathbf{x}') = k_1(x_i, \tilde{\mathbf{x}}, x'_i, \tilde{\mathbf{x}}') + k_2(x_j, \tilde{\mathbf{x}}, x'_j, \tilde{\mathbf{x}}'), \quad (33)$$

where $\tilde{\mathbf{x}}$ again represents $\mathbf{x}$ with x_i and x_j omitted. Under the competing model $\mathcal{M}_1$, joint causation is permissible, and so we permit the use of a kernel that can model the joint causation:

$$F|\mathcal{M}_1 \sim \mathcal{GP}(0, k_3). \quad (34)$$

The kernels k_1, k_2, k_3 are typically chosen from the same family, e.g., ARD-SE kernels. If priors over the hyperparameters are given, we may marginalize over the hyperparameter space and perform the model selection in a fully Bayesian manner. As an alternative, a hybrid approach in which we separately

optimize the hyperparameters for each kernel under each model from training data, and then perform the comparison using the 'best' hyperparameters for each model on test data. In the latter cases, we use the training data to obtain a posterior of the hyperparameters that is approximated by a Dirac delta function located at the best values of the hyperparameters. In either case, the posterior values $P(\mathcal{M}_q|\mathbf{X},\mathbf{y})$ are obtained via Bayes rule and modelling the data $\mathbf{X},\mathbf{y}$ by GPs based on the kernels of each model. When the prior over the models is uniform, the comparison of posterior simplifies to be a comparison of the model likelihoods [47].

When more specific models or hypotheses are of interest, the multiple hypothesis testing approach can be invoked to produce the corresponding direct tests, but still, a separate GPR model must be constructed for each test. Since the procedure of computing individual GPs and their likelihoods can also be parallelized, this approach can be implemented efficiently. However, the number of possible competing hypotheses grows exponentially with the dimension, so this approach is not scalable without modification.

A greedy approach to detect all the interactions is to repeatedly sift through each pair of features and to run the corresponding binary test. For each $i = 1,...,D$ and $j = i+1,...,D$, we assume that $F \sim \mathcal{GP}(0,k)$ for two possible cases: in the first case, we propose no change to the kernel k, and in the second case we modify k to ensure that $\partial^2 k/\partial x_i \partial x_j = 0$.[3] This approach would require $D(D-1)/2$ comparisons and two models to be learned per comparison, so it requires us to train $D(D-1)$ GPR models. However, as the complexity of the kernels increases, training the GPR models may potentially becomes prohibitive after many iterations. The number of comparisons could be reduced if we decide to proceed only with binary tests of the variables that produce $\partial_i \partial_j F$ close to zero.

E. Spectral Approximations to GPR

For a large number of training points N, using a GPR model as presented becomes prohibitive since the evaluation of the likelihood function requires the inversion of an $N \times N$ matrix. There are several approximations to GPR that are readily available [25], [48], [49], and the corresponding estimators of the MDCE can be derived for these approximations. In this section, we will focus on approximations obtained when one attempts to "sparsify" the spectral representation of the GP [50], [51], because they lend themselves to a rather straightforward analysis.

The motivation for sparse spectrum GPs, following from [50], is as follows: Consider *stationary* kernel function $k(\mathbf{x},\mathbf{x}')$, that is, a kernel function that depends only on $\mathbf{r} = \mathbf{x} - \mathbf{x}'$. Bochner's theorem [52] states that $k(\mathbf{r})$ can be expressed as the Fourier transform of a finite positive measure, i.e., a scaled probability measure $\sigma_0^2 p_k(\mathbf{v})$. This allows us to express the kernel in terms of an expectation,

$$\begin{aligned} k(\mathbf{x},\mathbf{x}') = k(\mathbf{r}) &= \int e^{2\pi i \mathbf{v}^\top \mathbf{r}} \sigma_0^2 p_k(\mathbf{v}) d\mathbf{v} \\ &= \sigma_0^2 \mathbb{E}_{p_k}\left(e^{2\pi i \mathbf{v}^\top \mathbf{r}}\right) \\ &= \sigma_0^2 \mathbb{E}_{p_k}\left(e^{2\pi i \mathbf{v}^\top \mathbf{x}} e^{-2\pi i \mathbf{v}^\top \mathbf{x}'}\right). \end{aligned} \tag{35}$$

The expression in (35) suggests that if one approximates the expectation using Monte Carlo integration, we may express the kernel through a series of samples $\mathbf{v}_m \sim p_k(\mathbf{v})$, $m = 1,...,M$. Since the power spectrum is symmetric about 0, we may also use $-\mathbf{v}_m$ as samples, which allows us to cancel the imaginary parts of the expression and yield an estimator,

$$k(\mathbf{x},\mathbf{x}') \approx \frac{\sigma_0^2}{M} \sum_{m=1}^{M} \cos(2\pi \mathbf{v}_m^\top (\mathbf{x} - \mathbf{x}')). \tag{36}$$

Once the frequency vectors $\mathbf{v}_m$ in (36) have been sampled, we can approximate the GPR posterior mean (16) using trigonometric functions:

$$\begin{aligned} \mathbb{E}(F(\mathbf{x})|\mathbf{X},\mathbf{y},\mathbf{x}) &= \sum_{n=1}^{N} k(\mathbf{x},\mathbf{x}_n)\,\alpha_n \\ &\approx \hat{F}(\mathbf{x}) = \sum_{m=1}^{M} \alpha_m \sin(\mathbf{v}_m^\top \mathbf{x}) + \beta_m \cos(\mathbf{v}_m^\top \mathbf{x}). \end{aligned} \tag{37}$$

We can extend the sparse spectrum GP approach to estimate the MDCE by differentiating the expression in (37) and obtain

$$\frac{\partial^2 \hat{F}(\mathbf{x})}{\partial x_i \partial x_j} = -\sum_{m=1}^{M} \alpha_m v_{m,i} v_{m,j} \sin(\mathbf{v}_m^\top \mathbf{x}) \tag{38}$$

$$- \sum_{m=1}^{M} \beta_m v_{m,i} v_{m,j} \cos(\mathbf{v}_m^\top \mathbf{x}). \tag{39}$$

In Section V, we compare the performance of sparse spectrum GPR to exact GPR when estimating the MDCE.

V. Experiments

We now explore various aspects of the MDCE approach through several examples and experiments. MATLAB code to reproduce these results is available online[4]. We include the following examples:

A. Comparison of different GP kernels for estimating the MDCE.
B. A time series example that compares MDCE with Volterra modeling. Additionally, we compare both methods to the Bayes detector.
C. MDCE estimation in a system with "linear confounders."
D. Comparison of exact GPR and sparse GPR.
E. A real data experiment using housing data from New Taipei City.

[3] By symmetry of the covariance function, this also means that $\partial^2 k/\partial x_i' \partial x_j' = 0$.

[4] See https://github.com/KurtButler/joint_causation.

Function	SNR	Kernel SE	ARD-SE	Additive	Matérn 5/2	Periodic
Local interaction	5dB	**1.082**	1.480	1.556	1.293	2.202
Local interaction	20dB	**0.252**	0.258	**0.252**	0.382	1.144
Egg box	5dB	0.377	0.385	0.385	0.844	**0.128**
Egg box	20dB	0.023	0.024	0.024	0.108	**0.004**
Egg box + Bilinear	5dB	7.053	7.250	7.234	**6.625**	17.179
Egg box + Bilinear	20dB	**0.656**	0.713	0.710	0.971	12.473

TABLE I: Mean-square-errors (MSE) of the GP mean as an estimator of the MDCE, for various functions with two inputs. The signal-to-noise ratio (SNR) is also shown for each example.

A. Comparison of kernels

To compare the performance of multiple kernels, we considered the general problem estimating the MDCE of x_1, x_2 on y in a nonlinear additive noise model:

$$x_1, x_2 \sim \mathcal{U}(-2, 2) \tag{40}$$
$$\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2) \tag{41}$$
$$y := F(x_1, x_2) + \varepsilon. \tag{42}$$

The noise variance σ_ε^2 was selected such that the resulting signal-to-noise ratio (SNR), defined to be $\mathbb{E}(F^2)/\sigma_\varepsilon^2$, was either 5dB or 20dB (we considered both cases). To try a variety of examples, we used three functions F:

1) A "local interaction" function,

$$F(x_1, x_2) = \sin(x_2) + \cos(2x_1)\cos(3x_2)\sigma(x_1), \tag{43}$$

where $\sigma(x) \triangleq 1/(1 + \exp(-5x))$ is a sigmoid function. Due to the presence of the sigmoid function, interactions in the model are only significant when $x_1 > 1$. We show the local interaction function and its MDCE function in Fig. 1.

2) An "egg box" function,

$$F(x_1, x_2) = \sin(2x_1)\sin(2x_2).$$

This function displays periodic behavior in both arguments, and has a spatially-varying MDCE function.

3) A sum of the egg box function and a bilinear term,

$$F(x_1, x_2) = \sin(2x_1)\sin(2x_2) + 4x_1x_2.$$

This function modifies the previous example to enforce an MDCE with nonzero mean. Additionally, the bilinear term has the potential to mask the influence of the eggbox function due to the difference in magnitude between the two summands.

In Fig. 2, we compare the estimates of the local interaction function with various kernels: the SE kernel, ARD-SE kernel, the additive kernel, Matérn 5/2 and the periodic kernel. After observing 300 samples of the original function immersed in white Gaussian noise (SNR = 20dB), we compare the various estimates of the MDCE. In practice, we might use any combination of the given kernels to improve our estimates, but we show each kernel separately to get a sense of the properties of each individual kernel. We also show the predictive mean-square-error (MSE) for each case, measured as the averaged difference between F and $\hat{F}$ averaged (numerically) across the grid.

In Table I, we systematically compare the results for each function and SNR combination, averaged across 100 randomly generated data sets.

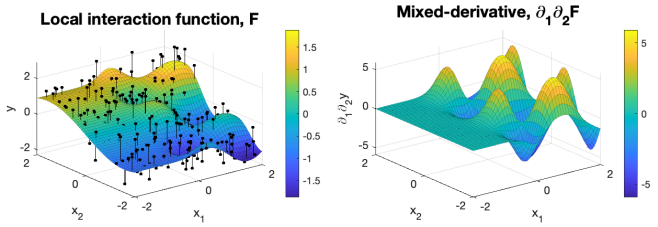

Fig. 1: Local interaction function $F(x_1, x_2)$, as given in (43), and its corresponding MDCE $\partial_1\partial_2 F$, visualized as functions of x_1 and x_2. Observations of the function immersed in white Gaussian noise are shown as black dots. The signal-to-noise ratio is 20dB. Given noisy observations of F, the goal is to use GPR to model $\partial_1\partial_2 F$, which is unobserved.

B. Time series example

To demonstrate our approach on a time series example, we consider the Volterra model used in nonlinear system identification [17], [18]. In the second-order *bilinear Volterra model*, the output signal y_t is modeled as a bilinear function of another signal x_t and its lags,

$$\begin{aligned} y_t &:= F(x_t, x_{t-1}, \cdots, x_{t-T}) + w_t \\ &= \sum_{i=0}^{T} a_i x_{t-i} + \sum_{i=0}^{T}\sum_{j=0}^{T} b_{ij} x_{t-i}x_{t-j} + w_t, \end{aligned} \tag{44}$$

where w_t is white Gaussian noise. The variance of w_t was chosen such that the signal-to-noise ratio, defined as the power ratio of the signal $s_t := F(x_t, \cdots, x_{t-T})$ to w_t, is 20 dB. In our simulation, we let x_t be an autoregressive process,

$$x_t = 0.85x_{t-1} + v_t, \tag{45}$$

where $v_t \sim \mathcal{N}(0, 1)$. To pick coefficients for the Volterra model, we sampled $a_i \sim \mathcal{N}(0, 1)$ independently and we fixed $b_{ij} \in \{0, \pm 1\}$. In Fig. 3, we show a realization of this process.

In Fig. 4, we compare the Hessian matrix $(\mathbf{H}_F)_{ij} = \partial_i\partial_j F$ of the model obtained to the averaged MDCE, the coefficients of a fitted Volterra model with bilinear kernel, and the Bayes detector. To estimate the averaged MDCE using GPR, we averaged the estimate of each $\partial_i\partial_j F$ over the sampled points. The GPR estimates of the averaged MDCE are comparable to directly fitting a Volterra model, but GPR did not require us

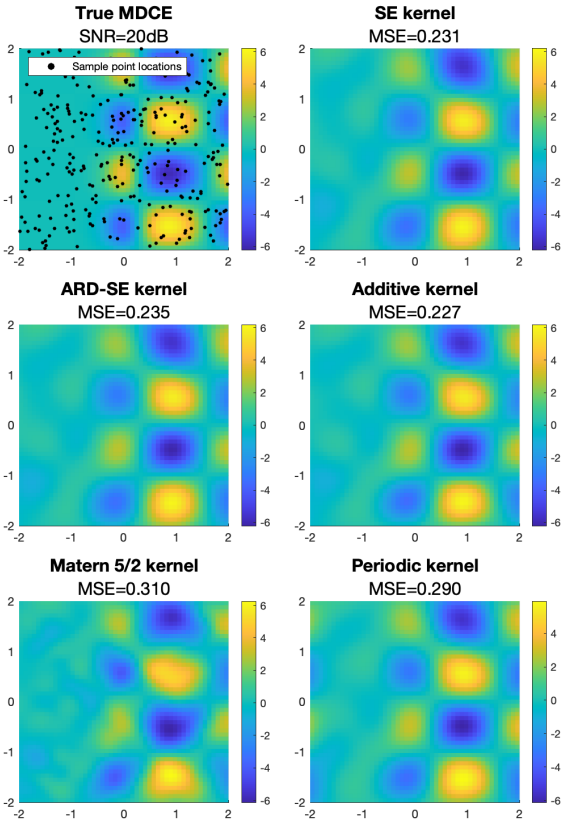

Fig. 2: We use heatmaps to visualize how different kernels produce different estimates of the MDCE, given the observations of the local interaction function in (43). The posterior means of the MDCE are shown for each choice of kernel. All kernels demonstrated the ability to interpolate within the region in which data were observed, but the additive kernel yielded the lowest fitting error.

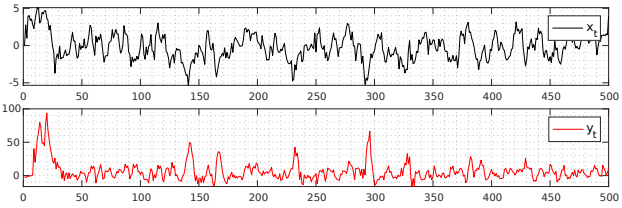

Fig. 3: A realization of the Volterra model in (44). The time series x_t is generated as an autoregressive process, via (45). The signal y_t is then generated as a function of x_t, according to (44). Given only observations of x_t and y_t, we then study the input-output relationship using the methods in Fig. 4.

to assume the relationship $x_t \to y_t$ to have a bilinear form. The Bayes detector, with a uniform prior over models, is also evaluated, and we found that it detected the location of the nonzero interactions fully. Combining the Bayes detector with the MDCE regression estimates can yield a better model of the Hessian matrix, with only weak assumptions.

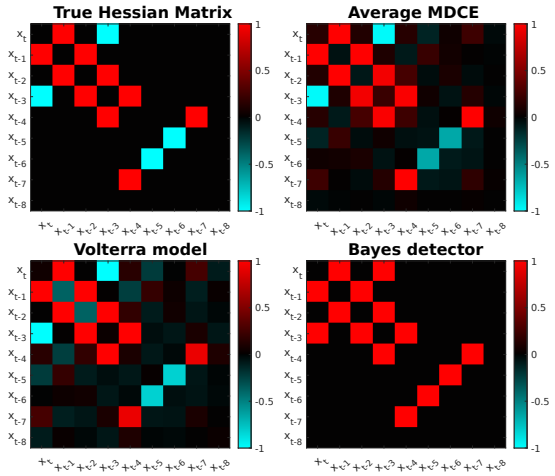

Fig. 4: Study of a Volterra system using MDCE. Time series x_t and y_t were generated as in Fig. 3. The Hessian matrix of the Volterra model, (44), is depicted, and is shown to be accurately estimated by the MDCE (averaged across the input space). The Hessian matrix is interpreted as a measure of joint causation between features. A parametric approach, using a bilinear Volterra model of the data, yields a similar pattern. Finally, we can also employ the Bayes detector, (31), to decide which entries of the Hessian are zero, which yields the correct joint causalities in this case.

C. *Linear confounders*

In this experiment, we consider a situation in which the MDCE estimator is robust to the presence of a certain class of common-cause confounder. Recalling equations (10) and (11), we consider the following probabilistic causal model with a confounder variable z:

$$z \sim \mathcal{N}(0, \sigma_z^2), \tag{46}$$

$$\mathbf{x} := \mathbf{a}z + \mathbf{w}_x, \tag{47}$$

$$y := (\mathbf{x}^\top \mathbf{B}\mathbf{x})(1 + \beta z) + \mathbf{1}^\top \mathbf{x} + 5z + w_y, \tag{48}$$

where σ_z^2 is the variance of the confounder z, the constants $\mathbf{a}, \mathbf{B}$ are fixed, and $\mathbf{w}_x \sim \mathcal{N}(\mathbf{0}, \sigma_x^2 \mathbf{I})$ and $w_y \sim \mathcal{N}(0, \sigma_y^2)$ are independent Gaussian noises. The parameter β is a control parameter that modulates how z can influence the MDCE. In particular, consider that under this model the MDCE of x_i and x_j on y is given by

$$\frac{\partial^2 y}{\partial x_i \partial x_j} = (B_{ij} + B_{ji})(1 + \beta z),$$

where B_{ij} are the entries of the matrix $\mathbf{B}$. When $\beta = 0$, the confounder z is 'linear' and does not influence the MDCE, and as a result we anticipate that one will be able to accurately the MDCE when the signal-to-noise (SNR) ratio[5] is reasonable. When $\beta \neq 0$, we expect that the MDCE estimates will be distorted depending on the strength of the confounder, in particular, depending on the magnitude of σ_z.

[5]Here, we are considering the signal to be $\mathbf{x}^\top \mathbf{B}\mathbf{x} + \mathbf{1}^\top \mathbf{x}$ and the noise to be $5z + w_y$. If the SNR is too low, we cannot obtain clean information about the $\mathbf{x} \to y$ relationship. However for moderate values of the SNR, we should be able to learn a model $\mathbf{x} \to y$ that preserves the joint causalities.

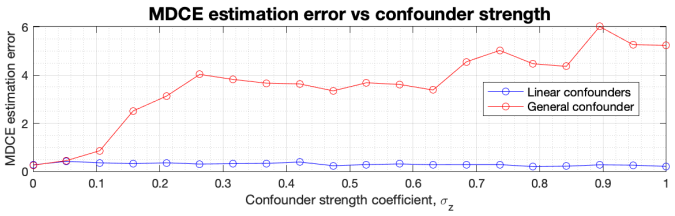

Fig. 5: Demonstration of the MDCE's robustness to linear confounders. We plot the median estimation error across repeated trials, where in each trial we compute the mean-square error of the MDCE estimate for a randomly generated data set. as a As the strength of the confounder, σ_z, increases, the MDCE accuracy should theoretically degrade. However, in the case that the confounder is 'linear,' and does not affect the second derivatives of the relationship, then the MDCE estimator is significantly more robust to the presence of the confounder.

In Figure 5, we consider estimation of the MDCE in the cases of $\beta = 0$ and $\beta = 0.5$. Additionally, in each case, we vary σ_z from 0 to 1. For each value of β and σ_z, we performed 300 trials, where in each trial we generate a random data set, and then we estimate the corresponding MDCEs. The parameter matrices $\mathbf{a}$ and $\mathbf{B}$ were randomly selected in each trial, where $\mathbf{a}$ had values sampled from $\mathcal{N}(0,1)$ and $\mathbf{B}$ had entries sampled from $\mathbf{N}(0,1)$, with an additional thresholding operation to induce sparsity in $\mathbf{B}$. We use $\dim(\mathbf{x}) = 3$ in all the experiments.

Figure 5 shows that for systems of the form of (48), two different paradigms of behavior are possible depending on β. In the 'linear confouder' scenario, when $\beta = 0$, the magnitude of the confounder does not significantly affect our ability to estimate the MDCE. This is contrast to the general situation in which any common cause confounder can negatively impact the study of causal strength. Thus, the MDCE as a measure of joint causation is robust to certain types of confounders.

D. Sparse GPR

We consider the following model to evaluate the sparse GP estimate of the MDCE:

$$\begin{aligned} x_1, x_2 &\sim \mathcal{U}(-2,2), \\ y &:= F(x_1,x_2) + \varepsilon, \end{aligned} \tag{49}$$

where

$$\begin{aligned} F(x_1,x_2) &= \sin(2x_1 + 2x_2)x_2 + 3\cos(x_1) + \sin(x_2), \\ \varepsilon &\sim \mathcal{N}(0,\sigma_\varepsilon^2), \end{aligned}$$

and σ_ε^2 is selected so that the SNR is 20dB.

In Fig. 6, we compare estimates of the function and the MDCE using both exact GPR, and sparse GPR approximations. The base GP model assumed an SE kernel. In this case, the SE kernel hyperparameters were selected to be $\ell = 0.9275$ and $\sigma_f = 2.6226$. For the sparse GPR, we used $M = 200$ frequencies from the power spectral density of this kernel. Under these conditions, the exact and sparse GPRs were able to obtain nearly indistinguishable models of the original

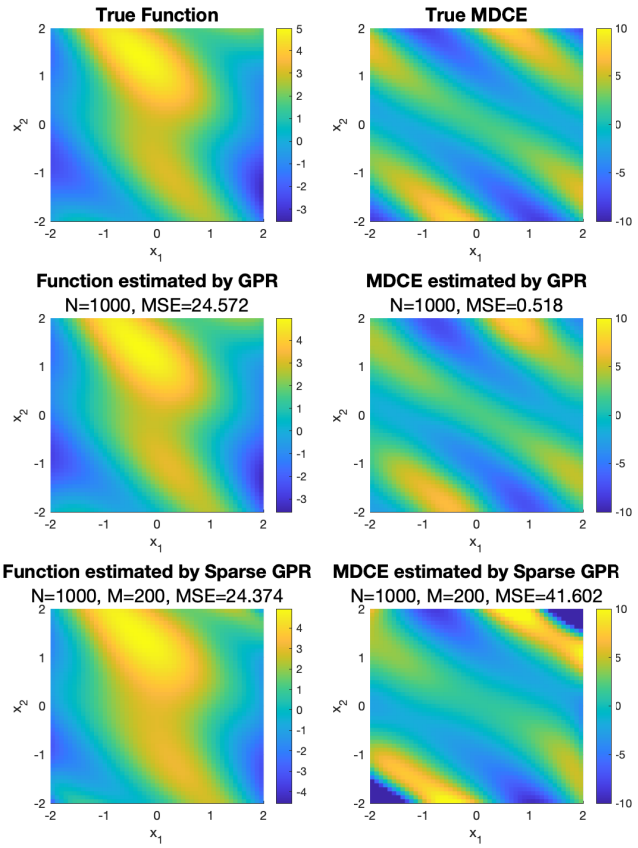

Fig. 6: Comparison of MDCE estimates using exact GPR and sparse approximation. The function F and its MDCE $\partial_1\partial_2 F$, as given in (49), are estimated using exact GPR and sparse GPR, respectively. The SNR in this example is 20dB. The exact GPR is able to accurately reproduce the function and its MDCE. The sparse approximation produces a reasonable estimate, but the error blows up near the boundary of the square.

function F. In comparison, the sparse GPR estimates of the MDCE were more erroneous near the boundaries of the square. However, within the interior of the square, we observed that the sparse GPR estimate of the MDCE was largely accurate. Naturally, standard GPR outperformed the sparse GPR using the same number of samples.

E. New Taipei City housing data

In our last example, we demonstrate how to analyze joint causation in a real data set. We used the New Taipei City housing data set from the UCI Machine Learning Repository [53]. This data set has 7 input features which are used to predict the price of a house in New Taipei City, and under the potential outcomes framework we can interpret this predictive model as a causal hypothesis. There are 414 samples in the data set, of which 364 were randomly selected to be used for training, and the remaining 52 were used to validate the model by checking the prediction error. We fit a GPR model with ARD-SE kernel to predict the house price y as a function of the input covariate vector $\mathbf{x}$. We cross-validated our model

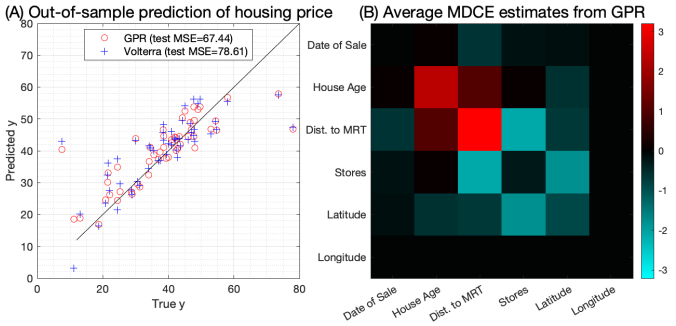

Fig. 7: Experimental results for the New Taipei City housing data set. **(A)** We show the out-of-sample prediction abilites for two models of the housing price: the first is a GPR model, and the second is a Volterra model using a bilinear Volterra kernel. The GPR model yielded a lower out-of-sample prediction mean-square-error (MSE). **(B)** The average MDCE under the GPR model, $\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),F\sim p(F|\mathbf{X},\mathbf{y})}(\partial_i\partial_j F(\mathbf{x}))$, after averaging over both the input location $\mathbf{x}$ and the model posterior uncertainty $p(F|\mathbf{X},\mathbf{y})$.

by randomizing the data set several times, confirming that our MDCE predictions are consistent as we change the training data.

In Figure 7, we visualize our model. The normalized prediction MSE is 0.252, so the GPR is able to explain the majority of the variance in the housing price using the given features. Since the model proved itself to be predictive, and it passed cross-validation, we find it appropriate to interrogate it with MDCE analysis. When viewing the MDCE, we can see that House Age and Distance to the MRT (metro) are both positive interacting features, which indicates that a house with both features together make a house perceived as more valuable than either feature does alone. Feature pairs with negative average MDCEs, like (Number of) Convenience Stores and Distance to the MRT, indicate that evaluators of the house are making tradeoffs when they assess the house's value. The ARD-SE kernel assigned a very low importance to the Longitude feature, so it was not important to the predictions, and this is reflected in the lack of joint causalities between Longitude and the other features.

VI. Conclusion

In this paper, we proposed the mixed differential causal effect as a measure of the strength of joint causation. By introducing Gaussian processes to model the posterior distribution over the mixed derivatives, we gained the ability to model the joint causation non-parametrically, making minimal assumptions about the underlying function. We applied the method to both static and time-series data and obtained results consistent with the existing theory, but also suggestive of future work and applications. The concept of joint causation is of interest to researchers in a number of fields including econometrics, climate science and medicine. The proposed tool for estimating joint causation can be used to understand studied phenomena without knowing the exact relationships between the causing and caused variables. Further, the tool can find applications in building novel parametric models that would allow for easier interpretations of acquired measurements.

Acknowledgment

The authors thank the National Science Foundation (Award 2212506) and the U.S. Department of Education for their support during the development of this work. The authors also thank the reviewers, whose feedback has greatly improved the content and quality of this paper, and Pei-Hsun Hsieh, for useful feedback during the writing of this work.

References

[1] Paul W Holland, "Statistics and Causal Inference," *Journal of the American Statistical Association*, vol. 81, no. 396, pp. 945–960, 1986.
[2] Sara Garofalo, Sara Giovagnoli, Matteo Orsoni, Francesca Starita, and Mariagrazia Benassi, "Interaction effect: Are you doing the right thing?," *PLoS One*, vol. 17, no. 7, pp. e0271668, 2022.
[3] Clive WJ Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.
[4] Donald B Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies.," *Journal of Educational Psychology*, vol. 66, no. 5, pp. 688, 1974.
[5] Judea Pearl, *Causality*, Cambridge University Press, 2009.
[6] Andrew T Court, "Measuring joint causation," *Journal of the American Statistical Association*, vol. 25, no. 171, pp. 245–254, 1930.
[7] Herman Wold, "Causality and econometrics," *Econometrica: Journal of the Econometric Society*, pp. 162–177, 1954.
[8] David R Cox, "Interaction," *International Statistical Review*, pp. 1–24, 1984.
[9] Sander Greenland, "Basic problems in interaction assessment.," *Environmental Health Perspectives*, vol. 101, no. suppl 4, pp. 59–66, 1993.
[10] John E Mathieu, Herman Aguinis, Steven A Culpepper, and Gilad Chen, "Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling.," *Journal of Applied Psychology*, vol. 97, no. 5, pp. 951, 2012.
[11] Hatice Ozer Balli and Bent E Sørensen, "Interaction effects in econometrics," *Empirical Economics*, vol. 45, pp. 583–603, 2013.
[12] Tyler J VanderWeele and Mirjam J Knol, "A tutorial on interaction," *Epidemiologic Methods*, vol. 3, no. 1, pp. 33–72, 2014.
[13] Tyler J VanderWeele, "On the distinction between interaction and effect modification," *Epidemiology*, pp. 863–871, 2009.
[14] Peter D Hoff, "Bilinear mixed-effects models for dyadic data," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 286–295, 2005.
[15] Chunrong Ai and Edward C Norton, "Interaction terms in logit and probit models," *Economics Letters*, vol. 80, no. 1, pp. 123–129, 2003.
[16] Jerome H Friedman and Bogdan E Popescu, "Predictive learning via rule ensembles," *The Annals of Applied Statistics*, pp. 916–954, 2008.
[17] Taiho Koh and E Powers, "Second-order Volterra filtering and its application to nonlinear aystem Identification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 6, pp. 1445–1455, 1985.
[18] Saban Ozer, Hasan Zorlu, and Selcuk Mete, "System identification application using Hammerstein model," *Sādhanā*, vol. 41, no. 6, pp. 597–605, 2016.
[19] Cecil Robinson and Randall E Schumacker, "Interaction effects: centering, variance inflation factor, and interpretation issues," *Multiple Linear Regression Viewpoints*, vol. 35, no. 1, pp. 6–11, 2009.
[20] Christophe Leys and Sandy Schumann, "A nonparametric method to analyze interactions: The adjusted rank transform test," *Journal of Experimental Social Psychology*, vol. 46, no. 4, pp. 684–688, 2010.
[21] Andrew Gelman, "Analysis of variance—why it is more important than ever," *The Annals of Statistics*, vol. 33, no. 1, feb 2005.
[22] Ilya M Sobol, "Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates," *Mathematics and Computers in Simulation*, vol. 55, no. 1-3, pp. 271–280, 2001.
[23] Art B Owen, "Sobol'indices and Shapley value," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 2, no. 1, pp. 245–251, 2014.
[24] Qiuling Yang, Mario Coutino, Geert Leus, and Georgios B Giannakis, "Autoregressive graph Volterra models and applications," *EURASIP Journal on Advances in Signal Processing*, vol. 2023, no. 1, pp. 1–21, 2023.

[25] Carl Edward Rasmussen and Christopher K Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.

[26] David K Duvenaud, Hannes Nickisch, and Carl Rasmussen, "Additive Gaussian processes," *Advances in Neural Information Processing Systems*, vol. 24, 2011.

[27] Francis Bach, "Exploring large feature spaces with hierarchical multiple kernel learning," *Advances in Neural Information Processing Systems*, vol. 21, 2008.

[28] Vladimir Vapnik, *Statistical Learning Theory*, Adaptive and Learning Systems for Signal Processing, Communications and Control. Wiley, 1998.

[29] Steven L Bressler and Anil K Seth, "Wiener–Granger causality: a well established methodology," *Neuroimage*, vol. 58, no. 2, pp. 323–329, 2011.

[30] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*, The MIT Press, 2017.

[31] B Schölkopf, D Janzing, J Peters, E Sgouritsa, K Zhang, and J Mooij, "On causal and anticausal learning," in *29th International Conference on Machine Learning (ICML 2012)*. International Machine Learning Society, 2012, pp. 1255–1262.

[32] Saber Salehkaleybar, AmirEmad Ghassami, Negar Kiyavash, and Kun Zhang, "Learning linear non-Gaussian causal models in the presence of latent variables," *Journal of Machine Learning Research*, vol. 21, no. 39, pp. 1–24, 2020.

[33] Yuhao Liu, Chen Cui, Daniel Waxman, Kurt Butler, and Petar M. Djurić, "Detecting confounders in multivariate time series using strength of causation," in *2023 31st European Signal Processing Conference (EUSIPCO)*. IEEE, 2023.

[34] Yixin Wang and David M Blei, "The blessings of multiple causes," *Journal of the American Statistical Association*, vol. 114, no. 528, pp. 1574–1596, 2019.

[35] Diederik P Kingma and Max Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[36] Clive WJ Granger, "Testing for causality: A personal viewpoint," *Journal of Economic Dynamics and Control*, vol. 2, pp. 329–352, 1980.

[37] Kurt Butler, Guanchao Feng, and Petar M. Djurić, "A differential measure of the strength of causation," *IEEE Signal Processing Letters*, 2022.

[38] Jerrold Marsden and Alan Weinstein, *Calculus III*, Springer, second edition, 1985.

[39] Eric Schulz, Maarten Speekenbrink, and Andreas Krause, "A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions," *Journal of Mathematical Psychology*, vol. 85, pp. 1–16, 2018.

[40] James Lloyd, David Duvenaud, Roger Grosse, Joshua Tenenbaum, and Zoubin Ghahramani, "Automatic construction and natural-language description of nonparametric regression models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2014, vol. 28.

[41] Petter Abrahamsen, *A review of Gaussian random fields and correlation functions*, Norsk Regnesentral/Norwegian Computing Center Oslo, 1997.

[42] Ercan Solak, Roderick Murray-Smith, WE Leithead, D Leith, and Carl Rasmussen, "Derivative observations in Gaussian process models of dynamic systems," *Advances in Neural Information Processing Systems*, vol. 15, 2002.

[43] Atilim Gunes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind, "Automatic differentiation in machine learning: a survey," *Journal of Marchine Learning Research*, vol. 18, pp. 1–43, 2018.

[44] Radford M Neal, *Bayesian Learning for Neural Networks*, vol. 118 of *Lecture Notes in Statistics*, Springer, 1996.

[45] David J.C. MacKay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, 2003.

[46] Bertil Matérn, *Spatial variation*, vol. 36, Springer Science & Business Media, 2013.

[47] Steven M. Kay, *Detection Theory*, vol. 2 of *Fundamentals in Statistical Signal Processing*, Prentice Hall, 1998.

[48] Qin Lu, Georgios V Karanikolas, and Georgios B Giannakis, "Incremental ensemble Gaussian processes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[49] Ali Rahimi and Benjamin Recht, "Random features for large-scale kernel machines," *Advances in Neural Information Processing Systems*, vol. 20, 2007.

[50] Miguel Lázaro-Gredilla, Joaquin Quinonero-Candela, Carl Edward Rasmussen, and Aníbal R Figueiras-Vidal, "Sparse spectrum Gaussian process regression," *The Journal of Machine Learning Research*, vol. 11, pp. 1865–1881, 2010.

[51] James Hensman, Nicolas Durrande, Arno Solin, et al., "Variational Fourier features for Gaussian processes," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 5537–5588, 2017.

[52] Michael L Stein, *Interpolation of Spatial Data: Some Theory for Kriging*, Springer Science & Business Media, 1999.

[53] I-Cheng Yeh, "Real Estate Valuation," UCI Machine Learning Repository, 2018, DOI: https://doi.org/10.24432/C5J30W.