A Differential Measure of the Strength of Causation

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

***Abstract*—The ability to quantify the strength of an interaction between events represented by random variables is important in many applications such as medicine and environmental science. We present the problem of measuring the strength of a causal interaction, starting from the linear perspective and generalizing to a nonlinear measure of causal influence, using a differential calculus approach. The proposed measure of causal strength is interpretable and may be estimated efficiently using Gaussian process regression. We validate our estimation approach on several synthesized data sets, considering both static variables and time series.**

***Index Terms*—causality, Gaussian processes, nonlinear systems, Simpson's paradox**

I. Introduction

In scientific problems, we are frequently interested in understanding the cause and effect relationships between events represented by the signals or variables observed in an experiment. Many techniques have been developed to detect the existence of causal interactions [1], [2], [3], but we often desire to quantify the intensity or strength of the interaction [4], [5], [6]. A good notion of causal strength enables many useful analyses, such as permitting doctors to determine the risk or severity associated with a physiological state or allowing climate scientists to rank several factors that cause global warming by their strength of causation.

Depending on the context, there are several measures of causal strength that may be useful. In our context, we are interested in studying a continuous random variable y whose value is determined by a set of continuous random variables $x_1, x_2, \cdots, x_D$. If the dependence of y on its causes is linear, e.g.

$$y := a_1 x_1 + \cdots + a_D x_D + \varepsilon, \tag{1}$$

where ε is a noise variable, then we may interpret the linear model coefficient a_i as a measure of the sensitivity of y to a small change in the value of x_i. The sensitivity of the effect variable to small changes in a cause variable will be called the causal strength or *causal effect* throughout this letter. We point out that the causal strength has a physical unit defined by the units of y and x_i, respectively.

The linear model shown in (1) is useful, but there are many phenomena that linear models cannot describe. The analogue of the linear coefficients for a nonlinear model are the partial derivatives. Generally, partial derivatives are not constants but rather functions that vary with the input variables. This feature complicates their interpretation, but also makes them more expressive. Approaches to measuring causation based on multivariate calculus have appeared before in some form in other disciplines [7], [8], and in [6] the idea of a nonlinear causal effect is briefly mentioned, but to the knowledge of the authors no data-driven estimator exists in the literature.

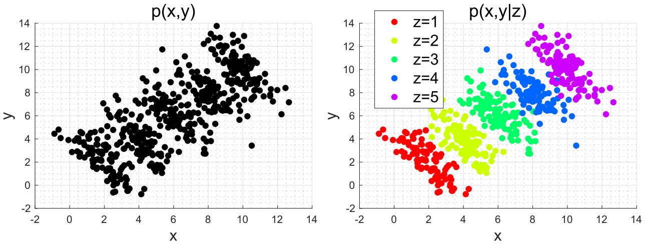

Fig. 1. Simpson's paradox [9] demonstrates that a measurement of causal effect depends critically on what features are included in the regression model. **Left:** A set of samples of points (x, y) from the generative model in (5). Fitting a regression model $y := ax + \varepsilon$ yields an ACE $a > 0$. **Right:** We show the same set of samples colored by the value of z. A joint regression model $y := ax + bz + \varepsilon$ now reveals $a < 0$.

The contribution of this letter is to propose a measure of causal strength based on partial derivatives and to provide a novel and nonparametric method of estimating the measure using Gaussian processes. The method is appropriate for any model in which the dependence between random variables is described by a differentiable function for any fixed realization.

We organize this letter as follows. In Section II, we propose the notion of differential causal effect as a measure of causal strength and show that it is a nonlinear generalization of average causal effect. We demonstrate how to estimate the proposed measure using Gaussian processes in Section III. We provide several examples and explain the connection of the method to causal inference in Section IV. We conclude this letter in Section V.

II. Background

The framework presented here can be used to study any relationship of the form

$$y := F(\mathbf{x}, \varepsilon), \tag{2}$$

where F is a differentiable function depending on a vector of inputs $\mathbf{x} = [x_1 \cdots x_D]^\top$ and a noise variable ε, and the symbol $:=$ denotes a causal assignment [6]. When the function F is linear, we may rewrite (2) as

$$y := a_1 x_1 + \cdots + a_D x_D + b\varepsilon. \tag{3}$$

The coefficient a_i controls how strong the coupling is between x_i and y, and when used as a measure of strength of causation it is called the *average causal effect* (ACE) of x_i on y

The authors thank the support of NSF under Award 2021002. Kurt Butler is supported by the U.S. Department of Education under the GAANN Fellowship.

The authors are with the Department of Electrical and Computer Engineering at Stony Brook University, Stony Brook, NY 11794-2350 USA.

[6]. Estimation of a_i can be done using linear regression or covariances since

$$\frac{\text{cov}(x_i, y)}{\text{var}(x_i)} = \frac{\sum_{j=1:D} a_j \text{cov}(x_i, x_j) + b\text{cov}(x_i, \varepsilon)}{\text{var}(x_i)} = a_i \quad (4)$$

whenever x_i is uncorrelated with the other inputs x_j and ε [6]. If the observed variables x_i are correlated, then linear regression can still be used to learn the ACE correctly. However, the problem can be much more challenging if there are unobserved variables. In [6], the ACE is defined using the *do*-operator [3], and in the supplemental material we comment on how to incorporate interventions into our analysis.

Simpson's paradox [9] demonstrates that the existence of unobserved variables can critically change the causal effect that we measure. We illustrate Simpson's paradox by considering the following generative model:

$$\begin{aligned} w_x, w_y &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \\ z &\overset{\text{i.i.d.}}{\sim} \mathcal{U}(\{1,...,5\}), \\ x &:= 2z + w_x, \\ y &:= 4z - x + w_y, \end{aligned} \quad (5)$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian distribution with mean μ and variance σ^2, and $\mathcal{U}(A)$ denotes the uniform distribution over a finite set A. In Figure 1, we draw 100 samples from this model and compare two models fitted to this data set. We find that knowledge of the common cause variable z will affect the detect sign and magnitude of causal effect of x on y. The dilemma presented by Simpson's paradox is difficult and well known within the causality literature [6], [9]. As a result, when we say the "causal effect of x on y," we mean the causal effect of x on y *in the assumed model*.[1]

ACE for linear models is interpretable and straightforward to estimate, but for nonlinear functions F we need to choose a particular generalization. A natural choice is to consider the partial derivatives, denoted either $\partial F/\partial x_i$ or $\partial y/\partial x_i$, which we will call the *differential causal effect* (DCE) of x_i on y. The DCE measures how small changes in x_i will change the affected variable y. When F is linear, the differential cause effect coincides with the linear coefficients. When F is nonlinear, the DCE varies with the current value of $\mathbf{x}$ and ε, which may be advantageous in modeling nonlinear systems.

To give a motivating example, in neuroscience there is a popular Bayesian modeling framework called dynamic causal modeling (DCM) [10]. In DCM one considers a *bilinear model*[2] of neuronal dynamics, which in our notation could be expressed as

$$y := F(\mathbf{x}, u, \varepsilon) = \mathbf{a}^\top \mathbf{x} + \mathbf{b}^\top \mathbf{x} u + \varepsilon, \quad (6)$$

where we interpret $\mathbf{x}$ as the current state of a neuron population and u as an exogenous influence to the population. Differentiating (6) yields

$$\frac{\partial F(\mathbf{x}, u, \varepsilon)}{\partial x_i} = a_i + b_i u, \quad (7)$$

and hence the causal coupling of x_i and y depends on the state of the exogenous variable u. Thus, the partial derivative function in (7) allows us to model *modulatory effects*, i.e., causal strength that is amplified or inhibited by other variables in the system.

To use DCE in practice, we have two problems. The first is the estimation of the DCE from a finite set of observed data. To this end, in the next section, we propose an approach based on Gaussian process regression. The second issue is how to appropriately summarize the DCE for general signals. We suggest several potentially useful techniques in Section IV, namely the use of histograms, averaging, and bilinear model coefficients.

III. Proposed solution

To estimate the DCE from data, we employ Gaussian process regression (GPR). GPR is a flexible, non-parametric and Bayesian approach to learning functions from data [11]. Given a vector of covariates $\mathbf{x} = (x_1, x_2, ..., x_D)$ and a target quantity $y = F(\mathbf{x})$, GPR estimates the function F by placing a Gaussian process prior over the space of possible functions $F : \mathbb{R}^D \to \mathbb{R}$, where we assume zero mean and the covariance between $F(\mathbf{x})$ and $F(\mathbf{x}')$ is controlled by a kernel function $k(\mathbf{x}, \mathbf{x}')$. After receiving data, the GPR estimate $\hat{F}$ of the function is given by the posterior mean and can be expressed in closed form, as derived in [11], by

$$\begin{aligned} \hat{F}(\mathbf{x}) &= \mathbb{E}(F(\mathbf{x}) | \mathbf{X}, \mathbf{y}, \mathbf{x}) \quad (8) \\ &= \mathbf{k}_*(\mathbf{x})^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y} \quad (9) \\ &= \sum_{n=1}^{N} k\left(\mathbf{x}, \mathbf{x}^{(n)}\right) \alpha_n, \quad (10) \end{aligned}$$

where $\mathbf{y}$ is a vector of training observations, $\mathbf{X} \in \mathbb{R}^{N \times D}$ is a matrix of training inputs $\mathbf{x}^{(n)}$, and $\mathbf{x}$ is the test point corresponding to the value $F(\mathbf{x})$ that we would like to predict, and α_n is the n-th entry in the vector $(\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y}$. Since the terms α_n do not depend on the particular test point $\mathbf{x}$, they are constants from the perspective of the derivative. Hence, the estimated DCE takes on a general form:

$$\frac{\partial \hat{F}}{\partial x_i} = \sum_{n=1}^{N} \frac{\partial k\left(\mathbf{x}, \mathbf{x}^{(n)}\right)}{\partial x_i} \alpha_n. \quad (11)$$

The derivative $\partial k/\partial x_i$ of the kernel depends on the particular choice of kernel used in GPR, but (11) clarifies how a change of kernel modifies our estimate of the DCE. For many popular kernels, we may compute $\partial k/\partial x_i$ easily. A common kernel is the squared-exponential (SE) kernel [11]:

$$k_{\text{SE}}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\sum_{j=1}^{D} \frac{(x_j - x_j')^2}{2\ell^2}\right), \quad (12)$$

where ℓ and σ_f are parameters. The derivative of $k(\mathbf{x}, \mathbf{x}^{(n)})$ shown in (11) is straightforward to compute, i.e.,

$$\frac{\partial k_{\text{SE}}\left(\mathbf{x}, \mathbf{x}^{(n)}\right)}{\partial x_i} = \sigma_f^2 \exp\left(-\sum_{j=1}^{D} \frac{(x_j - x_j^{(n)})^2}{2\ell^2}\right) \frac{x_i - x_i^{(n)}}{-\ell^2}. \quad (13)$$

[1] The problem here is that the inferred causal effect depends critically on what we regress on. As a result, we should always consider the causal strength as dependent on our current model [1].

[2] The bilinear models used in DCM are somewhat more complicated than the one in (6), but we simplify the model to exemplify modulated causation.

A popular extension of the SE kernel is the *SE kernel with automatic relevance detection* (ARD-SE) [12]. The ARD-SE kernel modifies (12) to let the parameter ℓ vary for each input dimension,

$$k_{\text{ARD-SE}}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\sum_{j=1}^{D} \frac{(x_j - x_j')^2}{2\ell_j^2}\right). \tag{14}$$

As a result, the ARD-SE kernel may automatically weight the importance of each input dimension x_i by varying the parameter ℓ_i. The corresponding modification to the derivative is also immediate

$$\frac{\partial k_{\text{ARD-SE}}\left(\mathbf{x}, \mathbf{x}^{(n)}\right)}{\partial x_i} = \sigma_f^2 \exp\left(-\sum_{j=1}^{D} \frac{(x_j - x_j^{(n)})^2}{2\ell_j^2}\right) \frac{x_i - x_i^{(n)}}{-\ell_i^2}. \tag{15}$$

Another common choice of kernel is the Matérn kernels. Matérn Gaussian processes are used because they can impose a restriction on the differentiability of the GP posterior samples [11]. The SE, ARD-SE and Matérn kernels are all universal kernels, meaning that they can be used to approximate any continuous function [13]. The Matérn 3/2 and 5/2 kernels are known to be once and twice continuously differentiable respectively, and they are given by the following equations:

$$k_{\text{Mat3/2}}(\mathbf{x}, \mathbf{x}^{(n)}) = \left(1 + \frac{\sqrt{3}r_n}{\ell}\right) \exp\left(-\frac{\sqrt{3}r_n}{\ell}\right), \tag{16}$$

$$k_{\text{Mat5/2}}(\mathbf{x}, \mathbf{x}^{(n)}) = \left(1 + \frac{\sqrt{5}r_n}{\ell} + \frac{5r_n^2}{3\ell^2}\right) \exp\left(-\frac{\sqrt{5}r_n}{\ell}\right), \tag{17}$$

where $r_n = ||\mathbf{x} - \mathbf{x}^{(n)}||$ is the Euclidean distance of the test point $\mathbf{x}$ from the n-th training point. The corresponding derivatives can again be computed,

$$\frac{\partial k_{\text{Mat3/2}}}{\partial x_i} = \frac{-3r_n}{\ell^2} \exp\left(-\frac{\sqrt{3}r_n}{\ell}\right) \frac{x_i - x_i^{(n)}}{r_n}, \tag{18}$$

$$\frac{\partial k_{\text{Mat5/2}}}{\partial x_i} = \left(\frac{-5r_n}{3\ell^2} - \frac{5\sqrt{5}r_n^2}{3\ell^3}\right) \exp\left(-\frac{\sqrt{5}r_n}{\ell}\right) \frac{x_i - x_i^{(n)}}{r_n}. \tag{19}$$

Naturally there are many other candidate kernels that we could discuss, but computation of their derivatives is completely analogous to the examples shown here. Interesting discussion on kernel selection and design can be found in [11], [14].

In large datasets, the direct implementation of GPR is prohibited by the difficultly in inverting the resulting large covariance matrix. In these cases it is common to approximate the kernel to improve scalability [15]. In such cases, derivations like above can be employed to produce estimators of the DCE, but this is beyond the scope of the current work.

IV. Results

To validate our approach, we study three simple models. Each model is intended to highlight different aspects of the DCE approach, and the utility of GPR as a tool for estimating the DCE in a data-driven manner.

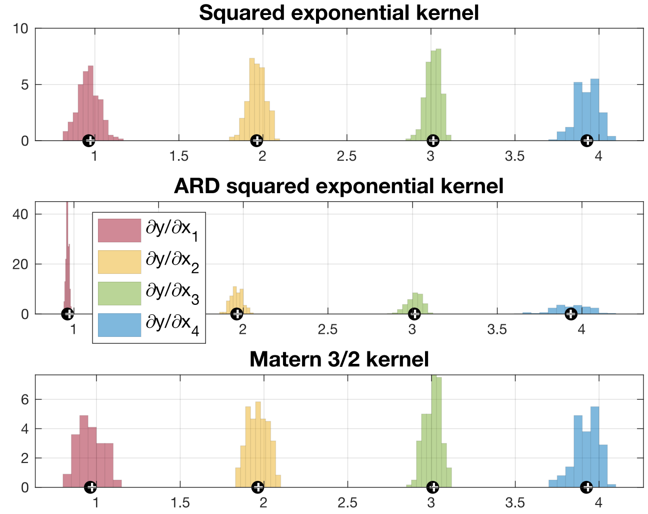

Fig. 2. DCE recovers the linear coefficients for the linear model in (20). Using three different kernels, the DCE estimated from 200 data points averages on the correct values. The black dots represent the mean values of each distribution, and the white crosses on top of each dot represents the corresponding estimate obtained using ordinary least squares.

A. Linear model

First, we demonstrate that the GPR-based estimate of DCE can reproduce the coefficients of a linear model. In Figure 2, we consider 200 data points sampled from the following model:

$$\begin{aligned} x_i &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \qquad i = 1, ..., 4, \\ \varepsilon &\sim \mathcal{N}(0,1), \\ y &:= x_1 + 2x_2 + 3x_3 + 4x_4 + \varepsilon. \end{aligned} \tag{20}$$

We estimate the differential causal estimate by regressing y on $x_1, ..., x_4$.

We observe that the average GPR-estimate coincides with the least squares model coefficients for each covariate and each kernel. However, we did not assume the functional form of the relationship in (20) during inference, which suggests that the average DCE will generally reproduce the ACE. We include in the supplemental material some results suggesting that this behavior is typical, even when the generative process is nonlinear.

B. Nonlinear model

Next, we consider a system in which x and y are related by a nonlinear function. We observe 500 samples from the following model:

$$\begin{aligned} x, z &\overset{\text{i.i.d.}}{\sim} \mathcal{U}(0,5), \\ \varepsilon &\sim \mathcal{N}(0,1), \\ f(x) &= \sin(x) + \cos(2x) + \sin(3x) + 0.1x^2, \\ y &:= f(x) + \cos(z) + \varepsilon. \end{aligned} \tag{21}$$

The DCE $\partial y/\partial x$ is only a function of x in this model.

In Figure 3, we show the result of regressing y on x using GPR with an SE kernel, and then we show that the estimated

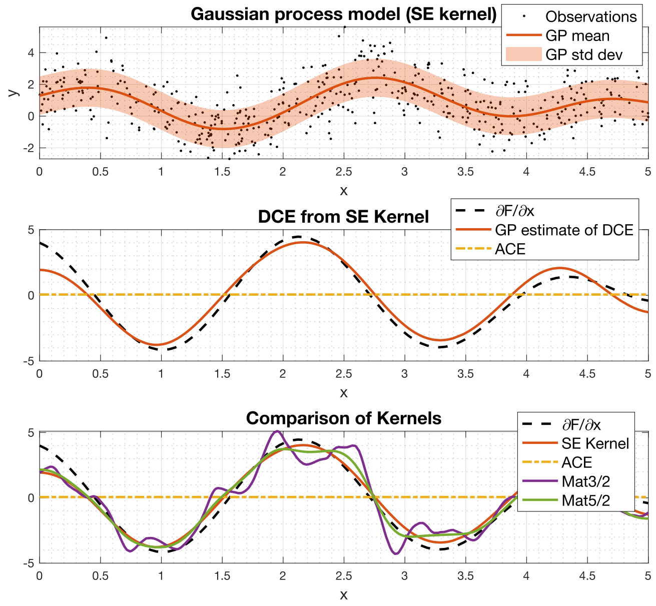

Fig. 3. DCE for the nonlinear model in (21). **Top:** The Gaussian process posterior mean and one standard deviation of variation for the data in the figure is shown, learned using the SE kernel. **Middle:** The DCE is estimated using SE kernel. The ACE is also shown, but is approximately zero. **Bottom:** Comparison of DCE estimates for several kernels.

DCE is close to the true causal effect as we vary the input x. Visualizing the ACE on the same plot, we see that it is close to zero (about -0.2), indicating that the ACE will not detect any causal influence from x to y because the distribution of DCE values has mean zero. However, the DCE is only zero in mean, and on average the magnitude of the DCE is about 2.1. Hence, the DCE encodes information about the causal strength that is more difficult to detect using the ACE.

Additionally in Figure 3, we compare the estimates of the DCE using the SE, Mat3/2 and Mat5/2 kernels. The ARD-SE kernel performed indistinguishably from the SE kernel, and so it was not plotted. We observe that while all kernels provided appropriate estimates, the Mat3/2 was more oscillatory than the other kernels. This is possibly due to the Mat3/2 kernel's limited differentiability, which suggests that kernel design may be important to refine a DCE estimator.

C. Modulated causation

As a final example, we consider the following signals:

$$\begin{aligned} w_x, w_y &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \\ x(t) &:= \sin\left(\frac{t}{2}\right) + \cos\left(\frac{\sqrt{2}t}{6}\right) + w_x, \\ z(t) &:= \frac{1}{1+\exp(15\sin(t/20))}, \\ y(t+1) &:= 0.9y(t) + (\beta + z(t))x(t) + w_y, \end{aligned} \tag{22}$$

where we selected $\beta = 0.1$. Note that $z(t)$ varies smoothly between 0 and 1, and $x(t)$ is just a superposition of sinusoids in noise.

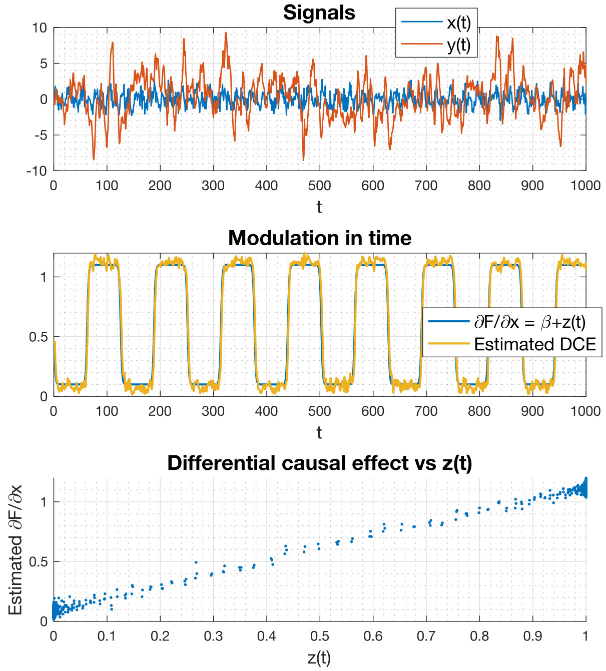

Fig. 4. DCE in a time series model. The signal y_t evolves according to $y_{t+1} = ay_t + (\beta + z_t)x_t + \varepsilon_t$, where z_t smoothly toggles between 0 and 1. **Top:** The signals $x(t), y(t)$ that are observed. **Middle:** The true DCE of $x(t)$ on $y(t+1)$ is $\partial y(t+1)/\partial x(t) = \beta + z(t)$. The GPR estimate tracks the true value closely in time. **Bottom:** We plot the estimated DCE against the corresponding values of $z(t)$. The strong linear trend suggests that a bilinear model may accurately capture the dynamics.

The DCE of $x(t)$ on $y(t+1)$ is given by $\beta+z(t)$. Hence, the next observation $y(t+1)$ depends strongly on both the current $x(t)$ only when the modulation signal $z(t)$ is high (near 1).

We observe 1000 samples from this model and attempt to estimate the DCE of $x(t)$ on $y(t+1)$ by regressing $y(t+1)$ on $x(t)$ and $y(t)$. In Figure 4, we show that our estimated DCE tracks the true modulation signal $z(t)$ in time. We see that DCE is approximately linear in $z(t)$, indicating that the underlying process is likely to be bilinear. Fitting a line to the scatterplot yields a slope of 1.007, close to the true value.

V. Conclusion

In this letter, we introduced a measure for strength of causation using ordinary calculus and presented a method to estimate it from data using Gaussian process regression. The proposed method was motivated by generalizing the analysis of linear models to nonlinear relationship. The proposed method can be used to study modulatory effects in neurological and biological systems.

References

[1] Clark Glymour, Kun Zhang, and Peter Spirtes, "Review of causal discovery methods based on graphical models," *Frontiers in Genetics*, vol. 10, pp. 524, 2019.

[2] Clive WJ Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[3] Judea Pearl, *Causality*, Cambridge University Press, second edition, 2009.

[4] Yonghan Jung, Shiva Kasiviswanathan, Jin Tian, Dominik Janzing, Patrick Blöbaum, and Elias Bareinboim, "On measuring causal contributions via do-interventions," in *International Conference on Machine Learning*. PMLR, 2022, pp. 10476–10501.

[5] Janine Witte, Leonard Henckel, Marloes H Maathuis, and Vanessa Didelez, "On efficient adjustment in causal graphs," *Journal of Machine Learning Research*, vol. 21, pp. 246, 2020.

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

[7] Ross M Stolzenberg, "The measurement and decomposition of causal effects in nonlinear and nonadditive models," *Sociological Methodology*, vol. 11, pp. 459–488, 1980.

[8] David Card, David S Lee, Zhuan Pei, and Andrea Weber, "Inference on causal effects in a generalized regression kink design," *Econometrica*, vol. 83, no. 6, pp. 2453–2483, 2015.

[9] Madelyn Glymour, Judea Pearl, and Nicholas P Jewell, *Causal Inference in Statistics: A Primer*, John Wiley & Sons, 2016.

[10] Karl J Friston, Lee Harrison, and Will Penny, "Dynamic causal modelling," *Neuroimage*, vol. 19, no. 4, pp. 1273–1302, 2003.

[11] Christopher K Williams and Carl Edward Rasmussen, *Gaussian processes for machine learning*, MIT Press, 2006.

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

[13] John Nicholson, Peter Kiessler, and D Andrew Brown, "A Kernel-Based Approach for Modelling Gaussian Processes with Functional Information," *arXiv preprint arXiv:2201.11023*, 2022.

[14] David Duvenaud, *Automatic model construction with Gaussian processes*, Ph.D. thesis, University of Cambridge, 2014.

[15] Qin Lu, Georgios Karanikolas, Yanning Shen, and Georgios B Giannakis, "Ensemble gaussian processes with spectral features for online interactive learning with scalability," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 1910–1920.