Scalar-Function Causal Discovery for Generating Causal Hypotheses with Observational Wearable Device Data

Valeriya Rogovchenko, Austin Sibu, and Yang Ni†
Department of Statistics, Texas A&M University,
College Station, TX 77843, USA
†*E-mail: yni@stat.tamu.edu*

Digital health technologies such as wearable devices have transformed health data analytics, providing continuous, high-resolution functional data on various health metrics, thereby opening new avenues for innovative research. In this work, we introduce a new approach for generating causal hypotheses for a pair of a continuous functional variable (e.g., physical activities recorded over time) and a binary scalar variable (e.g., mobility condition indicator). Our method goes beyond traditional association-focused approaches and has the potential to reveal the underlying causal mechanism. We theoretically show that the proposed scalar-function causal model is identifiable with observational data alone. Our identifiability theory justifies the use of a simple yet principled algorithm to discern the causal relationship by comparing the likelihood functions of competing causal hypotheses. The robustness and applicability of our method are demonstrated through simulation studies and a real-world application using wearable device data from the National Health and Nutrition Examination Survey.

Keywords: Causal identifiability, digital health, NHANES, observational data, wearable device.

1. Introduction

The rise of wearable devices has revolutionized the way we collect and analyze health data, offering an unprecedented wealth of information about human health and behavior. These devices such as accelerometers and continuous glucose monitors allow for frequent measurement of various variables over time including physical activities, sleep patterns, electrocardiogram signals, and blood glucose levels. The availability of these measurements enables researchers to ask questions that previously could not be answered, e.g., how to quantify the effect of physical activities on all-cause mortality? In these types of scenarios, often, one variable (e.g., physical activities) is longitudinal/functional and the other (e.g., mortality) is a scalar. Thus, many statistical methods such as scalar-on-function regression models[10,17] have been successfully deployed to estimate the association of the scalar-function pair.

The focus of this paper is, however, different from the existing literature for modeling wearable device data. Instead of association, we investigate whether it is possible to discern the *causal* relationship between a scalar and a function. More specifically, we aim to identify which of the scalar-function pair is more likely to be the cause or effect given observational data alone.

We introduce a novel *scalar-function causal discovery* method to generate data-driven causal hypotheses. Revealing the causality underlying observed data can deepen our understanding of the physical mechanism involved in the data-generating process and potentially pave the way for better health interventions and policy-making.

The field of causal discovery has seen a significant surge in interest and development over recent years due to wide-ranging applicability across various domains.[4,5,12–14,16,22] While traditional causal discovery methods are typically tailored to handle either continuous or discrete variables exclusively, real-world scenarios are often far more complex. For example, in the fields of social and health sciences, data frequently comprise a mix of different types of variables, necessitating more versatile approaches.

In such scenarios, one may either discard discrete data or convert continuous data into a discrete form;[9,20] either way, a lot of information contained in the original data is lost. In light of these limitations, there have been some recent developments to discover causality for mixed data.[19,21] However, these methods have only been developed for scalar variables, which cannot be used for functional data. To deal with functional data, some very recent works[6,23] have been proposed, which, however, cannot accomodate scalar and/or discrete data. In summary, to the best of our knowledge, there are no existing methods that can identify causality between a continuous functional variable and a binary scalar variable.

This paper, therefore, aims to fill this critical gap in the causal discovery literature so that digital health researchers will have a powerful tool to identify causality in a wide range of observational wearable device data. Our approach is based on a probabilistic causal model that quantifies the likelihood of each possible causal direction (from function to scalar or from scalar to function). We theoretically establish the causal identifiability property of our model under common causal assumptions. Equipped with the identifiability property, we can simply identify causal directions based on likelihood functions.

We conduct simulation studies to assess the empirical identifiability of the proposed method. In addition, to validate our method in real-world scenarios, we present an application with two variables that have a clear causal relationship. Specifically, we consider mobility conditions and physical activities. Since it is clear that mobility issues may lead to reduced activities, we will test whether our method can correctly identify such causal relationship without prior knowledge using the National Health and Nutrition Examination Survey (NHANES) data.

The rest of the paper is organized in the following way. In Section 2, we describe the proposed scalar-function causal discovery model, theoretically prove that the causal relationship is identifiable, and develop a likelihood-based estimation procedure. In Section 3, we evaluate the proposed method through various simulations as well as a real wearable device dataset from NHANES, demonstrating its capability to correctly identify the true causal relationship. We conclude our paper with a brief discussion in Section 4.

2. Method

2.1. *Notations*

We use capital letters to denote random variables and small letters to denote their realized values. We use boldface to denote vectors or matrices and non-boldface to denote scalars. With a slight abuse of notation, we use $P(\cdot)$ to denote both probability mass and density functions, which can be understood from the context as it is determined by the type of the random variable under consideration. Let $\mathbb{M}^{n\times n}$ be the cone of $n \times n$ positive definite matrices.

2.2. *Causal Probability Model*

We are interested in identifying the causal relationship between two statistically dependent random variables: a random binary variable $Y \in \{0, 1\}$ and a random function measured on n time points $\boldsymbol{X} = (X(t_1), ..., X(t_n))^\top \in \mathbb{R}^n$. One can view these functional measurements as a finite realization of an infinite stochastic process $X(\cdot)$ such as the Gaussian process.[18]

We consider two competing causal hypotheses[a],

$$H_0 : \boldsymbol{X} \to Y \text{ or } \boldsymbol{X} \text{ causes } Y$$

vs

$$H_1 : Y \to \boldsymbol{X} \text{ or } Y \text{ causes } \boldsymbol{X}$$

Under each hypothesis, we will set up a probability model. Specifically, let $P_{\boldsymbol{X}\to Y}(\boldsymbol{X} = \boldsymbol{x}, Y = y)$ denote the probability model of H_0 and $P_{Y\to\boldsymbol{X}}(\boldsymbol{X} = \boldsymbol{x}, Y = y)$ denote the probability model of H_1. Using the probability chain rule, we have

$$\begin{aligned} P_{\boldsymbol{X}\to Y}(\boldsymbol{X} = \boldsymbol{x}, Y = y) &= P_{\boldsymbol{X}\to Y}(Y = y \mid \boldsymbol{X} = \boldsymbol{x}) \cdot P_{\boldsymbol{X}\to Y}(\boldsymbol{X} = \boldsymbol{x}), \\ P_{Y\to\boldsymbol{X}}(\boldsymbol{X} = \boldsymbol{x}, Y = y) &= P_{Y\to\boldsymbol{X}}(\boldsymbol{X} = \boldsymbol{x} \mid Y = y) \cdot P_{Y\to\boldsymbol{X}}(Y = y), \end{aligned} \tag{1}$$

where $P_{\boldsymbol{X}\to Y}(Y = y \mid \boldsymbol{X} = \boldsymbol{x})$ and $P_{\boldsymbol{X}\to Y}(\boldsymbol{X} = \boldsymbol{x})$ are respectively the conditional and marginal probability distributions under $H_0 : \boldsymbol{X} \to Y$ and similarly $P_{Y\to\boldsymbol{X}}(\boldsymbol{X} = \boldsymbol{x} \mid Y = y)$ and $P_{Y\to\boldsymbol{X}}(Y = y)$ are those under $H_1 : Y \to \boldsymbol{X}$. Next, we will discuss the choice of these four probability distributions.

For the marginal distribution of Y, we assume it to be a Bernoulli distribution with success probability $\rho \in (0, 1)$,

$$P_{Y\to\boldsymbol{X}}(Y = y) = \rho^y(1 - \rho)^{1-y}. \tag{2}$$

For the marginal distribution of $\boldsymbol{X}$, we assume it to be a multivariate Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^n$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{M}^{n\times n}$,

$$P_{\boldsymbol{X}\to Y}(\boldsymbol{X} = \boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{3}$$

where $\mathcal{N}(\boldsymbol{x} \mid \cdot, \cdot)$ is the Gaussian probability density function evaluated at $\boldsymbol{x}$.

To model the conditional distribution of Y given $\boldsymbol{X}$, we adopt a linear logistic regression,

$$\log \frac{P_{\boldsymbol{X}\to Y}(Y = 1|\boldsymbol{X} = \boldsymbol{x})}{P_{\boldsymbol{X}\to Y}(Y = 0|\boldsymbol{X} = \boldsymbol{x})} = \alpha_0 + \boldsymbol{x}^\top\boldsymbol{\alpha}_1,$$

[a]Note that we are not performing null hypothesis testing. Our method is exploratory.

where α_0 is the intercept and $\boldsymbol{\alpha}_1 \neq \mathbf{0} \in \mathbb{R}^n$ are the slopes. That is, Y is conditionally Bernoulli,

$$P_{\boldsymbol{X}\to Y}(Y = y|\boldsymbol{X} = \boldsymbol{x}) = \phi_{\boldsymbol{x}}^y(1-\phi_{\boldsymbol{x}})^{1-y} \tag{4}$$

with the success probability depending on $\boldsymbol{X}$ through a sigmoid transformation,

$$\phi_{\boldsymbol{x}} = \frac{1}{1+e^{-\alpha_0-\boldsymbol{x}^\top\boldsymbol{\alpha}}}.$$

To specifiy $P_{Y\to\boldsymbol{X}}(\boldsymbol{X} \mid Y)$, we employ a multivariate linear regression model,

$$\boldsymbol{X} = \boldsymbol{\beta}_0 + Y\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^n$ are the intercepts, $\boldsymbol{\beta}_1 \neq \mathbf{0} \in \mathbb{R}^n$ are the slopes, and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ are Gaussian errors with mean zero and covariance $\boldsymbol{\Omega}$. The multivariate linear regression model above implies,

$$P_{Y\to\boldsymbol{X}}(\boldsymbol{X} = \boldsymbol{x}|Y = y) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\theta}_y, \boldsymbol{\Omega}) \tag{5}$$

with $\boldsymbol{\theta}_y = \boldsymbol{\beta}_0 + y\boldsymbol{\beta}_1$.

Putting (1)-(5) together, we have

$$\begin{aligned} P_{\boldsymbol{X}\to Y}(\boldsymbol{X} = \boldsymbol{x}, Y = y) &= \phi_{\boldsymbol{x}}^y(1-\phi_{\boldsymbol{x}})^{1-y}\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ P_{Y\to\boldsymbol{X}}(\boldsymbol{X} = \boldsymbol{x}, Y = y) &= \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\theta}_y, \boldsymbol{\Omega})\rho^y(1-\rho)^{1-y} \end{aligned} \tag{6}$$

2.3. *Causal Identifiability*

Since we only have access to observational data, the two competing causal hypotheses may not be identifiable, i.e., $P_{\boldsymbol{X}\to Y}(\boldsymbol{X} = \boldsymbol{x}, Y = y) = P_{Y\to\boldsymbol{X}}(\boldsymbol{X} = \boldsymbol{x}, Y = y)$ for all $\boldsymbol{x}$ and y. For example, if both $\boldsymbol{X}$ and Y are Gaussian, they are not identifiable. Consequently, even with an infinite amount of data, one cannot tell these two causal models apart – clearly an undesirable feature. Fortunately, we will show, both theoretically and empirically, that the proposed model is identifiable.

Definition 1 (Causal Identifiability). We say H_0 and H_1 are identifiable if one cannot find any values of $\{\alpha_0, \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$[b] and $\{\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\Omega}, \rho\}$[c] such that $P_{\boldsymbol{X}\to Y}(\boldsymbol{X} = \boldsymbol{x}, Y = y) = P_{Y\to\boldsymbol{X}}(\boldsymbol{X} = \boldsymbol{x}, Y = y)$ for all $\boldsymbol{x}$ and y.

Under the causal sufficiency assumption (i.e., there is no unmeasured confounder) commonly adopted in the literature,[2,4,11,13,15,22,23] we have the following identifiability theorem.

Theorem 1 (Causal Identifiability). *Assuming causal sufficiency, the causal hypotheses H_0 and H_1 are identifiable under model* (6).

Proof. We will show by contradiction. Suppose,

$$P_{\boldsymbol{X}\to Y}(\boldsymbol{X} = \boldsymbol{x}, Y = y \mid \alpha_0, \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = P_{Y\to\boldsymbol{X}}(\boldsymbol{X} = \boldsymbol{x}, Y = y \mid \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\Omega}, \rho) \tag{7}$$

[b]The parameters of $P_{\boldsymbol{X}\to Y}(\boldsymbol{X} = \boldsymbol{x}, Y = y)$
[c]The parameters of $P_{Y\to\boldsymbol{X}}(\boldsymbol{X} = \boldsymbol{x}, Y = y)$

for all $\boldsymbol{x} \in \mathbb{R}^n$ and $y \in \{0, 1\}$. Summing up both sides of (7) over y from 0 to 1, we have

$$\sum_{y=0}^{1} P_{\boldsymbol{X}\to Y}(Y = y \mid \boldsymbol{X} = \boldsymbol{x}, \alpha_0, \boldsymbol{\alpha}) P_{\boldsymbol{X}\to Y}(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{y=0}^{1} P_{Y\to\boldsymbol{X}}(\boldsymbol{X} = \boldsymbol{x} \mid Y = y, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\Omega}) P_{Y\to\boldsymbol{X}}(Y = y \mid \rho) \tag{8}$$

The left-hand side of (8) is given by

$$\begin{aligned}
&\sum_{y=0}^{1} P_{\boldsymbol{X}\to Y}(Y = y \mid \boldsymbol{X} = \boldsymbol{x}, \alpha_0, \boldsymbol{\alpha}) P_{\boldsymbol{X}\to Y}(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= P_{\boldsymbol{X}\to Y}(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \sum_{y=0}^{1} P_{\boldsymbol{X}\to Y}(Y = y \mid \boldsymbol{X} = \boldsymbol{x}, \alpha_0, \boldsymbol{\alpha}) \\
&= P_{\boldsymbol{X}\to Y}(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}),
\end{aligned} \tag{9}$$

where the second equality is due to the law of total probability.

The right-hand side of (8) is given by

$$\begin{aligned}
&\sum_{y=0}^{1} P_{Y\to\boldsymbol{X}}(\boldsymbol{X} = \boldsymbol{x} \mid Y = y, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\Omega}) P_{Y\to\boldsymbol{X}}(Y = y \mid \rho) \\
&= \rho \cdot P_{Y\to\boldsymbol{X}}(\boldsymbol{X} = \boldsymbol{x} \mid Y = 1, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\Omega}) + (1 - \rho) \cdot P_{Y\to\boldsymbol{X}}(\boldsymbol{X} = \boldsymbol{x} \mid Y = 0, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\Omega}) \\
&= \rho \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\beta_0} + \boldsymbol{\beta_1}, \boldsymbol{\Omega}) + (1 - \rho) \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\beta_0}, \boldsymbol{\Omega}).
\end{aligned} \tag{10}$$

Note that (9) is a Gaussian distribution whereas (10) is a mixture of Gaussian distribution. Therefore, for them to be equivalent, we must have $\rho = 0$, $\rho = 1$, or $\boldsymbol{\beta}_1 = \mathbf{0}$, which are degenerated cases (i.e., either Y is deterministically 0 or 1, or $\boldsymbol{X}$ and Y are independent).

□

Although our theorem relies on the causal sufficiency assumption, the experiments in Section 3.1.3 empirically show that the proposed method is relatively robust to the presence of unmeasured confounders.

2.4. *Estimation*

Theorem 1 establishes a property of the probability model and therefore is a population-level result. It implies that for a large enough sample size, one can correctly identify the correct causal hypothesis even with observational data alone. For a finite sample, our identifiability result paves the way for a simple, yet useful, causal discovery algorithm based on the maximum likelihood estimation (MLE). We aim to determine whether $\boldsymbol{X}$ causes Y or vice versa by quantifying the respective likelihoods. Therefore, when provided with a dataset of N subjects, $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$, we conclude $H_0 : \boldsymbol{X} \to Y$ if

$$\max_{\alpha_0,\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\Sigma}} \prod_{i=1}^{N} P_{\boldsymbol{X}\to Y}(\boldsymbol{X}=\boldsymbol{x}_i, Y=y_i \mid \alpha_0,\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\Sigma}) > \max_{\boldsymbol{\beta}_0,\boldsymbol{\beta}_1,\boldsymbol{\Omega},\rho} \prod_{i=1}^{N} P_{Y\to\boldsymbol{X}}(\boldsymbol{X}=\boldsymbol{x}_i, Y=y_i \mid \boldsymbol{\beta}_0,\boldsymbol{\beta}_1,\boldsymbol{\Omega},\rho),$$

and $H_1 : Y \to \boldsymbol{X}$ otherwise. Note that the two competing hypotheses have the same model complexity (i.e., the same number of parameters) and hence a model complexity penalty, which is typically needed for model selection, is not necessary here. The factorized form of the proposed model (6) allows us to separately find the MLE of each of its four components using existing standard techniques.

However, we note that in our motivating application, $\boldsymbol{X}$ is high-dimensional (n = 1,440). For better statistical and computational efficiency, we choose to reduce its dimensionality before finding the MLE. Specifically, the functional principal component analysis (FPCA) is used, which can reduce the functional data into a few uncorrelated functional principal components (FPCs) that explain the most variation among all the functional bases. We decompose the covariance function of a stochastic process $X(\cdot)$ as,

$$Cov(X(s), X(t)) = \sum_{k=1}^{\infty} \lambda_k \psi_k(s) \psi_k(t),$$

where λ_k's are the nonnegative eigenvalues in descending order and $\psi_k(\cdot)$'s are the corresponding orthogonal eigenfunctions. By the Karhunen-Loève theorem,

$$X(t) = \mu(t) + \sum_{k=1}^{\infty} Z_k \psi_k(t),$$

where $\mu(t) = E[X(t)]$, $\{\psi_k(t)\}_{k=1}^{\infty}$ is referred to as the FPCs, and $\{Z_{ik}(t)\}_{k=1}^{\infty}$ denotes the corresponding FPC scores. In practice, we would choose the first $K \ll n$ FPC scores $\boldsymbol{Z} = (Z_1, \ldots, Z_K)^\top$ that explain 99% variance and replace $\boldsymbol{X}$ by $\boldsymbol{Z}$ in the proposed model when finding the MLEs.

Finally, to assess the uncertainty of our approach, we use the bootstrap[3] technique in our real data application. We first generate B bootstrap samples by resampling subjects with replacement. Each bootstrap sample has the same size as the original dataset. Then we apply our method to each bootstrap sample and record our choice between H_0 and H_1. The proportion of times that we choose H_0 or H_1 reflects our confidence toward each hypothesis.

3. Experiments

We first tested our model through various simulation scenarios on synthetic data where there is known ground truth. After confirming its effectiveness, we then applied our method to a real-world mobility-activity dataset, demonstrating its practical capability in generating plausible causal hypotheses.

3.1. *Simulations*

To assess the efficacy of the proposed model, we performed simulations on three different synthetic datasets including one with unmeasured confounders. Each simulation was repeated 500 times, measuring the accuracy by the frequency at which we correctly identified the true

hypothesis. By considering varying sample sizes N between 50 and 200, we investigated the asymptotic behavior of our method. Furthermore, we examined the performance of the model under various signal strengths. For ease of exposition, $\boldsymbol{\delta}_i$ always denote the standard Gaussian white noises hereafter, i.e., $\boldsymbol{\delta}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ where $\boldsymbol{I}$ is an identity matrix.

3.1.1. *Case 1: True Direction $\boldsymbol{X} \to Y$*

For each subject $i = 1, \ldots, N$, the functional data $\boldsymbol{x}_i$ were created by first sampling their mean $\boldsymbol{m}$ from a centered Gaussian process at $n = 30$ evenly spaced time points,

$$\boldsymbol{m} \sim \mathcal{GP}(0, \mathcal{K})$$

with the powered exponential covariance function,

$$\mathcal{K}(t, s) = \exp\{-|t - s|^{\kappa}\},$$

of which the power $\kappa = 1.9$, and then setting

$$\boldsymbol{x}_i = \boldsymbol{m} + \boldsymbol{\delta}_i.$$

We performed the FPCA[24] on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ using the R package `fdapace`, and retained first K FPCs that explained 99% variance. We denote the standardized FPC scores by $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N$.

To create the causal dependency of y_i on $\boldsymbol{x}_i$ through $\boldsymbol{z}_i$, we generated y_i from a probit regression,

$$y_i = \begin{cases} 1 & \text{if} \quad y_i^* > 0 \\ 0 & \text{if} \quad y_i^* \leq 0 \end{cases},$$

where

$$y_i^* = 0.5 + \boldsymbol{z}_i^{\top} \boldsymbol{\gamma} + \epsilon_i$$

with $\epsilon_i \sim \mathcal{N}(0, 1)$. Here $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)^{\top}$ is the direct causal effect (signal), which will be varied at three levels: weak ($\gamma_k = \pm 1$), moderate ($\gamma_k = \pm 1.5$), and strong ($\gamma_k = \pm 3$).

The simulation results are reported in Table 1, showing an expected trend: the stronger the signal is, the more accurately the true causal direction can be discerned. Also, the accuracy approaches 100% as the sample size increases for the moderate and strong signal cases. Even with the weak signal, the accuracy was still good, around 90%. Note that for a non-identifiability model, the expected accuracy is 50%.

3.1.2. *Case 2: True Direction $Y \to \boldsymbol{X}$*

Exploring the reverse causal direction, we first generated the binary cause variable y_i from a Bernoulli distribution with a success probability of 0.5. Then we generated the functional effect variable,

$$\boldsymbol{x}_i = \boldsymbol{m}_{y_i} + \boldsymbol{\delta}_i,$$

where $\boldsymbol{m}_y \sim \mathcal{GP}(0, \mathcal{K}_y)$ for $y = 0, 1$ with the powered exponential covariance function $\mathcal{K}_y$ of which the power κ depends on y. Specifically, $\kappa = 1.9$ if $y = 1$, and $\kappa = 0.3$ (strong signal), 1.1 (moderate signal), or 1.7 (weak signal) if $y = 0$.

Table 1: Simulations. Accuracy of the proposed model in determining true causal directions in synthetic datasets over 500 simulations.

Case	Confounder	Signal	Sample size			
			50	100	150	200
$\boldsymbol{X} \to Y$	None	*weak*	92.8%	88.8%	90%	87.8%
		moderate	92.8%	97.2%	97.6%	99.6%
		strong	93.8%	98.6%	99%	99.8%
	Functional		94.6%	98.2%	99.2%	99.2%
	Binary		92.4%	99.2%	100%	100%
$Y \to \boldsymbol{X}$	None	*weak*	82.2%	89%	89.4%	93.6%
		moderate	83.4%	90.2%	96%	97.2%
		strong	97.8%	99%	99.4%	99.8%
	Functional		39.8%	61.6%	75%	83.8%
	Binary		63.6%	77%	85.2%	85.2%

As anticipated, our simulation results (Table 1) show that as the signal or the sample size increases, the accuracy approaches 100%.

3.1.3. *Case 3: Hidden Confounders*

The Simulation Cases 1&2 above demonstrate the validity of Theorem 1, i.e., causal directions can be identified even with observational data alone. We now empirically assess the robustness of the proposed method with respect to the violation of the causal sufficiency assumption, i.e., we test whether our method can still identify the correct causal direction in the presence of unmeasured confounders.

Our methodology hinges on determining the causality between two distinct types of variables, binary scalar and continuous functional. Thus, accordingly, we considered that the unobserved confounders, generically denoted by $\mathcal{C}$, are also either binary scalar or continuous functional. Consequently, we investigated four separate scenarios depicted in Fig. 1. We generated data from these four causal graphs and hid $\mathcal{C}$ from our method (i.e., only took $\boldsymbol{X}$ and Y as the inputs of our algorithm). As before, we recorded the frequency at which we correctly identified the causal direction between $\boldsymbol{X}$ and Y.

In Fig. 1 (a)&(b) where the confounder is binary, we generated the confounder c_i from a Bernoulli distribution with success probability 0.5. In Fig. 1 (a), the mean $\boldsymbol{m}_{c_i}$ of $\boldsymbol{x}_i$ was generated from a conditional Gaussian process $\boldsymbol{m}_c \sim \mathcal{GP}(0, \mathcal{K}_c)$ with the powered exponential covariance function $\mathcal{K}_c$ of which the power κ depends on c. Specifically, $\kappa = 1.9$ if $c = 1$ and $\kappa = 1.5$ if $c = 0$. Then as before, we set $\boldsymbol{x}_i = \boldsymbol{m}_{c_i} + \boldsymbol{\delta}_i$. Finally, we generated y_i from a probit regression model, $y_i = 1$ if $y_i^* > 0$ and $y_i = 0$ otherwise, where

$$y_i^* = 0.5 + 3 \cdot \boldsymbol{z}_i^\top \mathbf{1}_K + 3 \cdot c_i + \epsilon_i$$

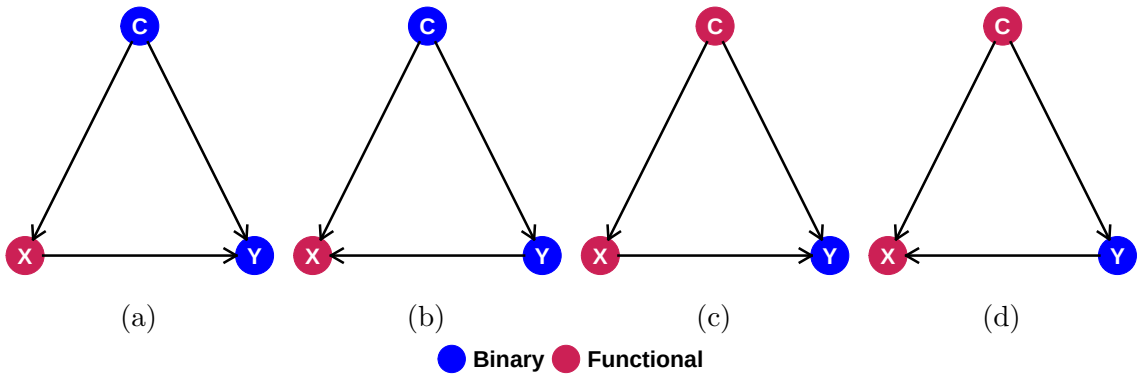

Fig. 1: Four confounding scenarios under consideration in Simulation Case 3.

with $\epsilon_i \sim \mathcal{N}(0,1)$, $\boldsymbol{z}_i$'s are the FPC scores of $\boldsymbol{x}_i$, and $\mathbf{1}_K = (1,\ldots,1)^\top$ is a vector of ones with length K.

In Fig. 1 (b), we generated y_i from a Bernoulli distribution with the success probability ρ_i dependent on c_i. More precisely, $\rho_i = 0.9$ if $c_i = 1$ and $\rho_i = 0.1$ if $c_i = 0$. Subsequently, the mean $\boldsymbol{m}_{c_i,y_i}$ of $\boldsymbol{x}_i$ was generated from a conditional Gaussian process $\boldsymbol{m}_{c,y} \sim \mathcal{GP}(0, \mathcal{K}_{c,y})$ with the powered exponential covariance function $\mathcal{K}_{c,y}$ of which the power κ depends on both c and y. To be specific, $\kappa = 1.9$ if $c = 1$ and $y = 1$, $\kappa = 0.5$ if $c = 1$ and $y = 0$, $\kappa = 1.0$ if $c = 0$ and $y = 1$, and $\kappa = 1.7$ if $c = 0$ and $y = 0$. Finally, we set $\boldsymbol{x}_i = \boldsymbol{m}_{c_i,y_i} + \boldsymbol{\delta}_i$.

In Fig. 1 (c)&(d), the functional confounder $\boldsymbol{c}_i$ was generated in the same way as $\boldsymbol{x}_i$ in Case 1 with $\kappa = 1.5$. Next, we performed the FPCA on $\boldsymbol{c}_1,\ldots,\boldsymbol{c}_N$ and retained the first J FPCs that explained 99% variance. We denote the standardized FPC scores by $\boldsymbol{d}_1,\ldots,\boldsymbol{d}_N$. In Fig. 1 (c), $\boldsymbol{m}$ was first generated from a centered Gaussian process $\boldsymbol{m} \sim \mathcal{GP}(0, \mathcal{K})$ with $\kappa = 1.9$. Then the dependence on the confounder was introduced by setting

$$\boldsymbol{x}_i = 0.5 + 5 \cdot \boldsymbol{m} + 5 \cdot \boldsymbol{c}_i + \boldsymbol{\delta}_i.$$

Finally, we generated y_i from a probit regression model, $y_i = 1$ if $y_i^* > 0$ and $y_i = 0$ otherwise, where

$$y_i^* = 0.5 + 5 \cdot \boldsymbol{z}_i^\top \mathbf{1}_K + 5 \cdot \boldsymbol{d}_i^\top \mathbf{1}_J + \epsilon_i$$

with $\epsilon_i \sim \mathcal{N}(0,1)$ and $\boldsymbol{z}_i$'s being the first K FPC scores of $\boldsymbol{x}_i$.

In Fig. 1 (d), we first generated y_i from a probit regression model, $y_i = 1$ if $y_i^* > 0$ and $y_i = 0$ otherwise, where

$$y_i^* = 0.5 + 3 \cdot \boldsymbol{d}_i^\top \mathbf{1}_J + \epsilon_i$$

with $\epsilon_i \sim \mathcal{N}(0,1)$. Then we generated mean processes $\boldsymbol{m}_{y_i}$ from a conditional Gaussian process $\boldsymbol{m}_y \sim \mathcal{GP}(0, \mathcal{K}_y)$. In this setting, the power κ of the powered exponential covariance function $\mathcal{K}_y$ depended on y: $\kappa = 1.9$ if $y = 1$, and $\kappa = 0.5$ if $y = 0$. To introduce the influence of the confounder, we defined

$$\boldsymbol{x}_i = \boldsymbol{m}_{y_i} + 3 \cdot \boldsymbol{c}_i + \boldsymbol{\delta}_i.$$

The results from these four confounding scenarios (Table 1) demonstrate the robustness of the proposed method. Particularly, as the sample size increased, our method achieved

increasingly better accuracy and was significantly better than a random guess for large sample sizes.

3.2. *Real Data*

Next, we applied the proposed methodology to the data collected by the NHANES. This extensive study, conducted by the Centers for Disease Control, gathered a wide range of health and nutritional information about the U.S. population, including sociodemographic characteristics and various health conditions. To demonstrate the utility of the proposed method, we are particularly interested in two variables, physical activities $\boldsymbol{X}$ captured via hip-attached accelerometers and an indicator variable of mobility issues Y derived from self-reported household interview data. Given the logical assumption of $Y \to \boldsymbol{X}$ in this scenario, we aim to verify if our method can correctly identify this causal direction, primarily seeking to validate the effectiveness of our method in accurately determining causation from a known truth.

3.2.1. *Data Preprocessing*

Utilizing the NHANES dataset, we accessed activity data from hip-worn accelerometers during the 2003–2004 and 2005–2006 study waves. The magnitude of acceleration (movement "intensity") was captured using the ActiGraph AM-7164, delivering an objective measure of physical activity and bypassing the inconsistencies of self-reported data. Participants were instructed to wear the device for seven consecutive days, excluding swimming and bathing periods. The raw data were segmented into one-minute intervals or "epochs" with intensity readings accumulated per epoch and saved in long format (each row is a subject-minute).

The well-formatted data are contained in the R package `rnhanesdata`.[7,8] Following the preprocessing procedure in their paper,[8] we included individuals aged 50 to 85 and omitted non-compliant individuals who have excessive missing accelerometer data, leaving us with $N = 3,198$ subjects.

The activity data for each individual were aggregated over the 7-day period and transformed via $\log(1+x)$. This dataset is organized in a $7N \times 1440$ matrix, with one row designated for each subject-day across all NHANES waves, where 7 denotes the days each subject wore the accelerometer, and 1440 corresponds to the total number of minutes in a day.

The presence of any mobility issues was represented as a binary variable, categorized as either "No difficulty" or "Any difficulty," based on responses from the Physical Functioning questionnaire. Individuals were classified under "Any difficulty" if they reported challenges in climbing 10 stairs, walking a quarter mile, abstained from these activities, or required special walking equipment. Overall, there are 32.4% subjects in the sample who experience any mobility of movement problem.

3.2.2. *Results*

We generated $B = 100$ bootstrap samples and successfully identified the correct causal direction across all samples from comparing the maximized likelihoods of $Y \to \boldsymbol{X}$ and $\boldsymbol{X} \to Y$: the mobility issue Y unambiguously impacts an individual's level of physical activity $\boldsymbol{X}$ with high

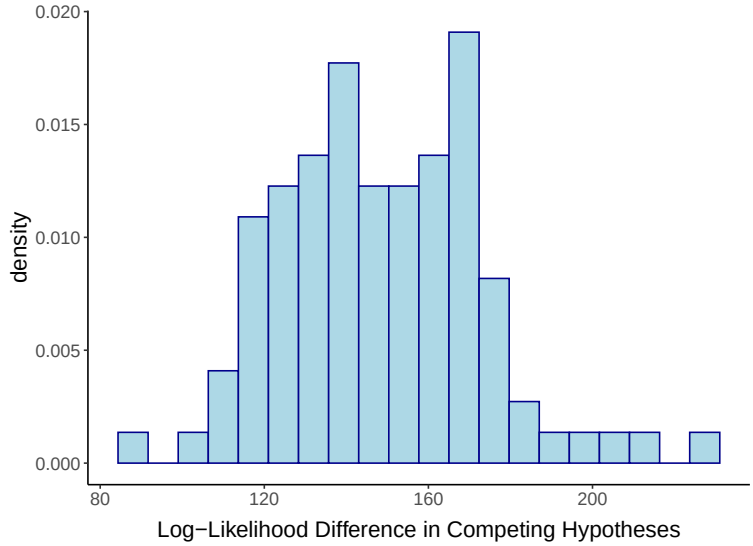

Fig. 2: Real data. Histogram depicting the maximized log-likelihood differences between the competing hypotheses $Y \to \boldsymbol{X}$ and $\boldsymbol{X} \to Y$.

confidence. As depicted in Fig. 2, the histogram illustrates the difference in the maximized log-likelihoods between these two competing hypotheses (the former minus the latter), which is noticeably bounded away from zero, meaning that $Y \to \boldsymbol{X}$ is far more likely than $\boldsymbol{X} \to Y$, which matches the presumed truth.

4. Discussion

In this paper, we have presented a new causal model for generating bivariate causal hypotheses with a continuous functional variable (e.g., physical activities) and a binary scalar variable (e.g., mobility issue indicator) in an exploratory fashion, which can provide insights as to which variable is more likely the cause. We theoretically proved that the underlying cause-effect relationship is identifiable with purely observational data under the causal sufficiency assumption. Empirically, we used a likelihood-based inference procedure and demonstrated the utility of the proposed method both under and beyond the causal sufficiency setting through simulation studies and a real-world wearable device application.

There are several areas where this paper could be strengthened and extended. First, our NHANES application has focused on physical activities and mobility issue because of their clear causal relationship. Having demonstrated it is possible to identify their causal relationship, we plan to analyze other variables in the data to generate causal hypotheses in an exploratory manner, which is an intended use of the proposed method.

Second, our identifiability theory operates under the assumption that there are no unmeasured confounders. Even though our empirical investigations have indicated a degree of robustness to the presence of confounders, a theoretical exploration of identifiability within this context would be interesting and particularly relevant in observational studies where the presence of unmeasured confounders is common.

Third, we have focused on the bivariate case and hence an extension to multivariate

cases and leveraging additional publicly available datasets can considerably broaden the method's applicability. For example, the brain electroencephalogram dataset[1] comprises electroencephalogram signals collected over various trials with distinct stimuli for two groups - alcoholics and controls. By viewing the electroencephalogram signals as multivariate functional data, a recent paper[23] attempts to discern the causal relationships among these functions. The multivariate extension of our method could potentially enrich this research by providing additional insights into the causal relationships modified by the experimental groups by treating the group as a binary variable. Moreover, it should be relatively straightforward to extend our method to incorporate multiple categorical scalar variables.

Finally, a Bayesian inference approach could be adopted especially for multivariate cases where efficient searching strategies in the causal graph space are required. A Bayesian approach would make it easier to make finite-sample inferences with natural uncertainty quantification for complex causal graphs.

Acknowledgment

Ni's research was partially supported by NIH 1R01GM148974-01 and NSF DMS-2112943.

References

1. H. Begleiter. EEG Database. UCI Machine Learning Repository, 1999. DOI: https://doi.org/10.24432/C5TS3D.
2. J. Choi, R. Chapkin, and Y. Ni. Bayesian causal structural learning with zero-inflated poisson Bayesian networks. *Advances in neural information processing systems*, 33:5887–5897, 2020.
3. B. Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer, 1992.
4. P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.
5. D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.
6. K.-Y. Lee and L. Li. Functional structural equation model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):600–629, 2022.
7. A. Leroux. *rnhanesdata: NHANES Accelerometry Data Pipeline*, 2023. R package version 1.02.
8. A. Leroux, J. Di, E. Smirnova, E. J. Mcguffey, Q. Cao, E. Bayatmokhtari, L. Tabacu, V. Zipunnikov, J. K. Urbanek, and C. Crainiceanu. Organizing and analyzing the activity data in nhanes. *Statistics in biosciences*, 11:262–287, 2019.
9. S. Monti and G. F. Cooper. A multivariate discretization method for learning bayesian networks from mixed data. *arXiv preprint arXiv:1301.7403*, 2013.
10. J. S. Morris. Functional regression. *Annual Review of Statistics and Its Application*, 2:321–359, 2015.
11. Y. Ni. Bivariate causal discovery for categorical data via classification with optimal label permutation. *Advances in Neural Information Processing Systems*, 35:10837–10848, 2022.
12. J. Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2):3, 2000.
13. S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

14. P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search, 2nd Edition*, volume 1 of *MIT Press Books*. The MIT Press, December 2001.
15. P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT press, 2000.
16. O. Stegle, D. Janzing, K. Zhang, J. M. Mooij, and B. Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. *Advances in neural information processing systems*, 23, 2010.
17. C. D. Tekwe, R. S. Zoh, M. Yang, R. J. Carroll, G. Honvoh, D. B. Allison, M. Benden, and L. Xue. Instrumental variable approach to estimating the scalar-on-function regression model with measurement error with application to energy expenditure assessment in childhood obesity. *Statistics in medicine*, 38(20):3764–3781, 2019.
18. J.-L. Wang, J.-M. Chiou, and H.-G. Müller. Functional data analysis. *Annual Review of Statistics and Its Application*, 3(1):257–295, 2016.
19. W. Wenjuan, F. Lu, and L. Chunchen. Mixed causal structure discovery with application to prescriptive pricing. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 5126–5134, 2018.
20. M. Yamayoshi, J. Tsuchida, and H. Yadohisa. An estimation of causal structure based on latent lingam for mixed data. *Behaviormetrika*, 47:105–121, 2020.
21. Y. Zeng, S. Shimizu, H. Matsui, and F. Sun. Causal discovery for linear mixed data. In B. Schölkopf, C. Uhler, and K. Zhang, editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 994–1009. PMLR, 11–13 Apr 2022.
22. K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. 647–655, 2009.
23. F. Zhou, K. He, K. Wang, Y. Xu, and Y. Ni. Functional bayesian networks for discovering causality from multivariate functional data. *arXiv preprint arXiv:2210.12832*, 2022.
24. Y. Zhou, S. Bhattacharjee, C. Carroll, Y. Chen, X. Dai, J. Fan, A. Gajardo, P. Z. Hadjipantelis, K. Han, H. Ji, C. Zhu, H.-G. Müller, and J.-L. Wang. *fdapace: Functional Data Analysis and Empirical Dynamics*, 2022. R package version 0.5.9.