Causal Modeling with Stochastic Confounders

Thanh Vinh Vo [1] **Pengfei Wei** [1] **Wicher Bergsma** [2] **Tze-Yun Leong** [1]

[1]School of Computing, National University of Singapore

[2]Department of Statistics, London School of Economics and Political Science

votv@comp.nus.edu.sg, dcsweip@nus.edu.sg, w.p.bergsma@lse.ac.uk, leongty@comp.nus.edu.sg

Abstract

This work extends causal inference in temporal models with stochastic confounders. We propose a new approach to variational estimation of causal inference based on a representer theorem with a random input space. We estimate causal effects involving latent confounders that may be interdependent and time-varying from sequential, repeated measurements in an observational study. Our approach extends current work that assumes independent, non-temporal latent confounders with potentially biased estimators. We introduce a simple yet elegant algorithm without parametric specification on model components. Our method avoids the need for expensive and careful parameterization in deploying complex models, such as deep neural networks in existing approaches, for causal inference and analysis. We demonstrate the effectiveness of our approach on various benchmark temporal datasets.

1 INTRODUCTION

Estimating the causal effects of an intervention or treatment on an outcome is an important problem in scientific investigations and real-world applications. For example: What is the effect of sleep-deprivation on health outcomes? How would family socio-economic status affect career prospects? What is the impact of a disease outbreak on the regional stock markets? Existence of *confounders* or confounding variables that affect both the treatment and the outcome may produce bias in the estimation, and hence complicate the causal inference process. Classical approaches to dealing with observed or visible confounders are the propensity score-based methods and their variants (Rubin, 2005). In cases with latent or hidden confounders, however, the treatment effects on the outcome cannot be directly estimated without further assumptions (Pearl, 2009; Louizos et al., 2017). For example, in crop planting, rainfall received in a field may affect both the fertilizer (since it may drain the fertilizer off the field) and the crop yield (through water the plants received). In this situation, *rainfall* is the hidden confounder as it may not be directly measurable in real life.

Recent studies in causal inference (Shalit et al., 2017; Louizos et al., 2017; Madras et al., 2019) mainly focus on static data, i.e., the observational data are time-independent and with independent and identically distributed (*iid*) noise. In many real-world applications, however, events change over time, e.g., each participant may receive an intervention multiple times and the timing of these interventions may differ across participants. In this case, the time-independent assumption does not hold, and causal inference would degenerate in the models that fail to capture the nature of time-dependent data. In practice, temporal confounders such as seasonality and long-term trends can potentially contribute to confounding bias. For example, the confounding soil fertility in crop planting may change over time, due to different reasons such as annual rainfall or soil erosion. Whenever soil fertility, which may or may not be directly measured, declines (or raises), it would possibly affect the *future* level of fertility. In real-life situations, there may be many such latent confounders that may not be interpretable. This motivates us to propose a causal modeling framework that captures these latent confounding variables that change over time, and a causal inference method that mitigates the stochastic confounding problem and reduces bias in estimating causal effects.

In this work, we introduce a framework to characterize the latent confounding effects over time in causal inference based on the structural causal model (SCM)

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

(Pearl, 2000). Inspired by recent work (e.g., Riegg, 2008; Louizos et al., 2017; Madras et al., 2019) that handle static, *independent* confounders, we relax this assumption by modeling the confounders as a stochastic process. This approach generalizes the *independent* setting to model confounders that have intricate patterns of interdependencies over time.

Many existing causal inference methods (e.g., Louizos et al., 2017; Shalit et al., 2017; Madras et al., 2019) exploit recent developments of deep neural networks. While effective, the performance of a neural network depends on many factors such as its structure (e.g., the number of layers, the number of nodes, or the activation function) or the optimization algorithm. Tuning a neural network is challenging; different conclusions may be drawn from different network settings. To overcome these challenges, we propose a nonparametric variational method to estimate the causal effects of interest using a kernel as a prior over the reproducing kernel Hilbert space (RKHS). Our main contributions are summarized as follows:

• We introduce a Causal Inference with Stochastic Confounders (CausalSC) method for temporal data that captures the interdependencies of confounders. This relaxes the independent confounders assumption in recent work (e.g., Louizos et al., 2017; Madras et al., 2019). Under this setting, we introduce the concepts of causal path effects and intervention effects, and derive approximation measures of these quantities.

• Our framework is robust and simple for accurately learning the relevant causal effects: given a time series, learning causal effects quantifies how an outcome is expected to change if we vary the treatment level or intensity. Our algorithm requires no information about how these variables are parametrically related, in contrast to the need for paramterizing a neural network.

• We develop a nonparametric variational estimator by exploiting the kernel trick in our temporal causal framework. This estimator has a major advantage: complex non-linear functions can be used to modulate the SCM with estimated parameters that turn out to have analytical solutions. We empirically demonstrate the effectiveness of the proposed estimator.

2 BACKGROUND AND RELATED WORK

The structural causal model (SCM) (Pearl, 1995, 2000) is a general causal modeling framework that builds on several seminal works, including structural equation models (Goldberger, 1972, 1973; Duncan, 1975, 2014), potential outcomes framework of Neyman (1923) and Rubin (1974), and graphical models (Pearl, 1988). The SCM consists of a triplet: a set of exogenous variables whose values are determined by factors outside the framework, a set of endogenous variables whose values are determined by factors within the framework, and a set of structural equations that express the value of each endogenous variable as a function of the values of the other (endogenous and exogenous) variables. Figure 1 (a) shows a causal graph with endogenous variables Y, Z, W. Here Y is the outcome variable, W is the intervention variable, and Z is the confounder variable. Exogenous variables are variables that are not affected by any other variables in the model, which are not explicitly in the graph. Causal inference evaluates the effects of an intervention on the outcome, i.e., $p(Y \mid \mathrm{do}(W = w))$, the distribution of the outcome Y induced by setting W to a specific value w. Our work focuses on the problem of estimating causal effects, which is different from the works of identifying causal structure (or causal discovery) (Spirtes et al., 2000), where several approaches have been proposed, e.g., Tian and Pearl (2001); Peters et al. (2013, 2014); Jabbari et al. (2017); Huang et al. (2019).

Estimators with unobserved confounders. The following efforts take into account the unobserved confounders in causal inference: Montgomery et al. (2000); Riegg (2008); de Luna et al. (2017); Kuroki and Pearl (2014); Louizos et al. (2017); Madras et al. (2019). Specifically, some proxy variables are introduced to replace or infer the latent confounders. For example, the household income of a student is a confounder that affects the ability to afford private tuition and hence the academic performance; it may be difficult to obtain income information directly, and proxy variables such as zip code, or education level are used instead. Figures 1 (b) and (c) present two causal graphs used by the recent causal inference algorithms in Louizos et al. (2017); Madras et al. (2019). The graphs contain latent confounder Z, proxy variable X, intervention W, outcome Y, and observed confounder S.

Estimators with observed confounders. Hill (2011); Shalit et al. (2017); Alaa and van der Schaar (2017); Yao et al. (2018); Yoon et al. (2018); Künzel et al. (2019); Oprescu et al. (2019); Nie and Wager (2020) follow the formalism of potential outcomes and these works do not take into account latent confounders but satisfy the strong ignorability assumption of Rosenbaum and Rubin (1983). Of interest is the inference mechanism by Alaa and van der Schaar (2017), where the authors model the counterfactuals as functions living in the reproducing kernel Hilbert space. In contrast, we work on a time series setting within a structural causal framework whose inference scheme directly benefits from the generalization of the empirical risk minimization framework.

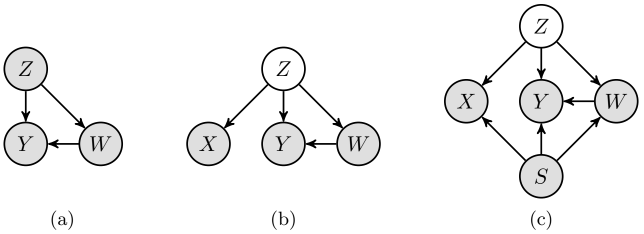

Figure 1: The SCM framework: approaches of modeling causality. (a) all variables are observed; (b) the confounder Z is latent and being inferred using proxy variable X (Louizos et al., 2017); (c) there is an additional observed confounder S (Madras et al., 2019).

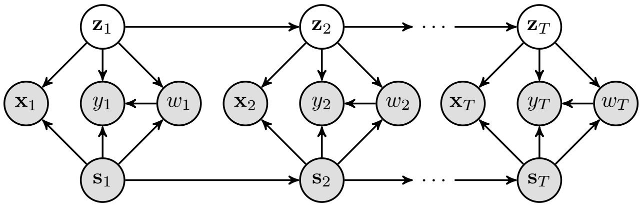

Figure 2: Causal inference with Stochastic Confounders (CausalSC) for temporal data.

Estimators for temporal data. Little attention has been paid to learning causal effects in non-*iid* setting (Guo et al., 2020; Bica et al., 2020b). Lu et al. (2018) formalize the Markov Decision Process under a causal perspective and the emphasis is mainly on sequential optimization. This work does not discuss the estimation of average treatment effects using observational data. Li and Bühlmann (2018) model potential outcomes for temporal data with observed confounders using state-space models. The models proposed are linear and quadratic regression. Ning et al. (2019) propose causal estimates for temporal data with observed confounders using linear models and developed Bayesian flavors inference methods. Bojinov and Shephard (2019) generalize the potential outcomes framework to temporal data and proposed an inference method with Horvitz–Thompson estimator (Horvitz and Thompson, 1952). This method, however, does not consider the existence of unobserved confounders. Bica et al. (2020a,b) formalize potential outcomes with observed and unobserved confounders to estimate counterfactual outcomes for treatment plans on each individual, where the outcomes are modelled with recurrent neural networks. The observed data of *each individual* in these two works are temporal and the individuals are independent with each other. In other words, there is a collection of observed time series for each individual. Our work, on the other hand, has no notion of individuals. We instead observe *one time series* for each feature of the data, i.e., the outcome, treatment, covariates and observed confounders (covariates and observed confounders may have several features). Thus, the problem setup of our work is different from that of Bica et al. (2020a,b). Several efforts (e.g., Kamiński et al., 2001; Eichler, 2005, 2007, 2009; Eichler and Didelez, 2010; Bahadori and Liu, 2012) analyse causation for temporal data based on the notion of Granger causality (Granger, 1980). These works focus on discovering causal structures, which is different from our work that estimates causal effects over a period of time.

3 OUR APPROACH

We introduce a Causal inference with Stochastic Confounders (CausalSC) framework based on SCM, as illustrated in Figure 2. Following Frees et al. (2004), we evaluate the causal effects within a *time interval* denoted by t. We assume that the interval is large enough to cover the effects of the treatment on the outcome. This assumption is practical as many interval-censored datasets are recorded monthly or yearly.

The latent confounder z. In real world applications, capturing all the potential confounders is not feasible as some important confounders might not be observable. When there are unobserved or latent confounders, causal inference from observational data is even more challenging and, as discussed earlier, can result in a biased estimation. The increasing availability of large and rich datasets, however, enables proxy variables for unobserved confounders to be inferred from other observed and correlated variables. In practice, the exact nature and structure of the hidden or latent confounder $\mathbf{z}$ are unknown. We assume the structural equation of the latent confounder at time interval t as follows:

$$\mathbf{z}_t = f_z(\mathbf{z}_{t-1}) + \boldsymbol{e}_t, \tag{1}$$

where the exogenous variable $\boldsymbol{e}_t \sim \mathsf{N}(\mathbf{0}, \sigma_z^2 \mathbf{I}_{d_z})$, the function $f_z \colon \mathcal{Z} \mapsto \mathcal{F}_z$ with $\mathcal{Z}$ is the set containing $\mathbf{z}_t$ and $\mathcal{F}_z$ is a Hilbert space, σ_z is a hyperparameter and $\mathbf{I}_{d_z}$ denotes the identity matrix. The choice of this structural equation is reasonable because each dimension of $\mathbf{z}_t$ maps to a real value which gives a wide range of possible values for $\mathbf{z}_t$. The function $f_z(\cdot)$ would control the effects of confounder at time interval $t-1$ to its future value at time interval t. With reference to an earlier example, the soil fertility in crop planting can be considered as confounder and this quantity changes over time. When the fertility of soil at time interval $t-1$ declines, possibly because of soil erosion, the fertility of soil at time interval t may also declines. Furthermore, the Gaussian assumption imposed on exogenous variable $\boldsymbol{e}_t$ makes it computational tractable for subsequent calculations.

The observed variables y, w, x and s. The d_x–dimensional observed features at time interval t, is denoted by $\mathbf{x}_t \in \mathbb{R}^{d_x}$. Similarly, the treatment variable at time interval t is given by w_t. y_t and $\mathbf{s}_t$ denote the outcome and the observed confounder at time interval t, respectively. $\mathbf{z}_t$ denotes an unobserved confounder. Let $\mathcal{Y}$, $\mathcal{W}$, $\mathcal{S}$, $\mathcal{X}$ be sets containing y_t, w_t, $\mathbf{s}_t$, $\mathbf{x}_t$, respectively, and let $\mathcal{F}_y$, $\mathcal{F}_w$, $\mathcal{F}_s$, $\mathcal{F}_x$ be Hilbert spaces. Let $f_y\colon \mathcal{Z}\times\mathcal{W}\times\mathcal{S} \mapsto \mathcal{F}_y$, $f_w\colon \mathcal{Z}\times\mathcal{S}\mapsto\mathcal{F}_w$, $f_s\colon \mathcal{S}\mapsto\mathcal{F}_s$, $f_x\colon \mathcal{Z}\times\mathcal{S}\mapsto\mathcal{F}_x$. We postulate the following structural equations:

$$\mathbf{s}_t = f_s(\mathbf{s}_{t-1}) + \boldsymbol{o}_t, \quad y_t = f_y(w_t, \mathbf{z}_t, \mathbf{s}_t) + v_t, \tag{2}$$

$$\mathbf{x}_t = f_x(\mathbf{z}_t, \mathbf{s}_t) + \boldsymbol{r}_t, \quad w_t = \mathbb{1}(\varphi(f_w(\mathbf{z}_t, \mathbf{s}_t)) \geq u_t), \tag{3}$$

where the exogenous variables $\boldsymbol{o}_t \sim \mathsf{N}(\mathbf{0}, \sigma_s^2\mathbf{I}_{d_s})$, $\boldsymbol{r}_t \sim \mathsf{N}(\mathbf{0}, \sigma_x^2\mathbf{I}_{d_x})$, $v_t \sim \mathsf{N}(0, \sigma_y^2)$ and $u_t \sim \mathsf{U}[0,1]$, $\mathbb{1}(\cdot)$ denotes the indicator function and $\varphi(\cdot)$ is the logistic function. The last structural equation implies that w_t is Bernoulli distributed given $\mathbf{z}_t$ and $\mathbf{s}_t$, and $p(w_t = 1|\mathbf{z}_t, \mathbf{s}_t) = \varphi(f_w(\mathbf{z}_t, \mathbf{s}_t))$. Similar reasoning holds for these assumptions in that they afford computational tractability in terms of time series formulation. Our model assumes that the treatment and outcome at time interval $t-1$ are independent of the treatment and outcome at time interval t given the confounders. These assumptions are important in real-life, for example, in the context of agriculture, the crop yield (outcome) of the current crop season probably does not affect crop yield in the next season, and similarly for the chosen fertilizer (treatment). However, they may be correlated through the latent confounders, e.g., soil fertility of the land.

Learnable functions. With Eqs. (1)-(3), we need to learn f_i, where $i \in \boldsymbol{\mathcal{A}} := \{y, w, z, x, s\}$. Standard methods to model these functions include linear models or multi-layered neural networks. Selecting models for these functions relies on many problem-specific aspects such as types of data (e.g., text, images), dataset size (e.g., hundreds, thousands, or millions of data points), and data dimensionality (high- or low-dimension). We propose to model these functions through an *augmented representer theorem*, to be desribed in Section 4.

3.1 Causal Quantities of Interest

Temporal data capture the evolution of the characteristics over time. Based on earlier work (Louizos et al., 2017; Pearl, 2009; Madras et al., 2019), we evaluate causal inference from temporal data by assuming that the confounders satisfy Eq. (1). The corresponding causal graph is shown in Figure 2. This relaxes the independent assumption in Louizos et al. (2017) and Madras et al. (2019). In particular, we aim to measure the causal effects of w_t on y_t given the covariate $\mathbf{x}_t$, where $\mathbf{x}_t$ serves as the proxy variable to infer latent confounder $\mathbf{z}_t$. This formulation subsumes earlier approaches (Louizos et al., 2017; Madras et al., 2019). Considering multiple time intervals, we further denote the vector notations for T time intervals as follows: $\mathbf{y} = [y_1,..., y_T]^\top, \mathbf{w} = [w_1,..., w_T]^\top, \mathbf{s} = [\mathbf{s}_1,..., \mathbf{s}_T]^\top, \mathbf{x} = [\mathbf{x}_1,..., \mathbf{x}_T]^\top, \mathbf{z} = [\mathbf{z}_1,..., \mathbf{z}_T]^\top, \mathbf{z}_0 = \emptyset, \mathbf{s}_0 = \emptyset$. We define *fixed-time* causal effects as the causal effects at a time interval t and *range-level* causal effects as the average causal effects in a time range. The fixed-time and range-level causal effects can be estimated using Pearl's *do*-calculus (Pearl, 2009). We model the unobserved confounder processes using the latent variable $\mathbf{z}_t$, inferred for each observation $(\mathbf{x}_t, \mathbf{s}_t, w_t, y_t)$ at time interval t. The interval is assumed to be large enough to cover the effects of the treatment w_t on the outcome y_t. This assumption is practical as many interval-censored datasets are recorded monthly or yearly.

Definition 1. *Let* $\mathbf{w}_1$ *and* $\mathbf{w}_2$ *be two treatment paths. The treatment effect path (or effect path) of* $t \in [1, 2,..., T]$ *is defined as follows:*

$$\mathrm{EP} := E[\,\mathbf{y} \,|\, \mathrm{do}(\mathbf{w}=\mathbf{w}_1), \mathbf{x}] - E[\,\mathbf{y} \,|\, \mathrm{do}(\mathbf{w}=\mathbf{w}_2), \mathbf{x}]. \tag{4}$$

The effect path (EP) is a collection of causal effects at time interval $t \in [1, 2, \ldots, T]$.

Definition 2. *Let* EP *be the effect path satisfying Definition* 1. *Let* $\mathrm{EP}_t \in \mathrm{EP}$ *be the effect at time interval* t. *The average treatment effect (ATE) in* $[T_1, T_2]$ *is defined as follows:*

$$\mathrm{ATE} := \left(\textstyle\sum_{t=T_1}^{T_2} \mathrm{EP}_t\right) / (T_2 - T_1 + 1). \tag{5}$$

The average treatment effect (ATE) quantifies the effects of a treatment path $\mathbf{w}_1$ over an alternative treatment path $\mathbf{w}_2$. This work focuses on evaluating the causal effects with binary treatment. The key quantity in estimating the effect path and average treatment effects is $p(\mathbf{y} \,|\, \mathrm{do}(\mathbf{w}), \mathbf{x})$, which is the distribution of $\mathbf{y}$ given $\mathbf{x}$ after setting variable $\mathbf{w}$ by an intervention. Following Pearl's back-door adjustment (Pearl, 2009) and invoking the properties of *d*-separation, the causal effects of $\mathbf{w}$ to $\mathbf{y}$ given $\mathbf{x}$ with respect to the causal graph in Figure 2 is as follows:

$$p(\mathbf{y}|\mathrm{do}(\mathbf{w}), \mathbf{x}) = \int p(\mathbf{y}|\mathbf{w}, \mathbf{z}, \mathbf{s}) p(\mathbf{z}, \mathbf{s}|\mathbf{x}) d\mathbf{s} d\mathbf{z}. \tag{6}$$

The expression in Eq. (6) typically does not have an analytical solution when the distributions $p(\mathbf{y} \,|\, \mathbf{w}, \mathbf{z}, \mathbf{s})$, $p(\mathbf{z}, \mathbf{s} \,|\, \mathbf{x})$ are parameterized by more involved impulse functions, e.g., a nonlinear function such as multi-layered neural network. Thus, the empirical expectation of $\mathbf{y}$ is used as an approximation. To do so, we first draw samples of $\mathbf{z}$ and $\mathbf{s}$ from $p(\mathbf{z}, \mathbf{s}|\mathbf{x})$, and then substitute these samples to $p(\mathbf{y}|\mathbf{w}, \mathbf{z}, \mathbf{s})$ to draw samples of $\mathbf{y}$. The whole procedure is carried out using forward sampling techniques under specific forms of

$p(\mathbf{y}|\mathbf{w},\mathbf{z},\mathbf{s})$, $p(\mathbf{z}|\mathbf{x},\mathbf{s},\mathbf{w},\mathbf{y})$, $p(\mathbf{y}|\mathbf{s},\mathbf{x},\mathbf{w})$, $p(\mathbf{w}|\mathbf{s},\mathbf{x})$ and $p(\mathbf{s}|\mathbf{x})$. In the next section, we present approximations of these probability distributions.

4 ESTIMATING CAUSAL EFFECTS

Estimating causal effects requires systematic sampling from the following distributions: $p(\mathbf{s}|\mathbf{x})$, $p(\mathbf{w}|\mathbf{s},\mathbf{x})$, $p(\mathbf{y}|\mathbf{s},\mathbf{x},\mathbf{w})$, $p(\mathbf{z}|\mathbf{x},\mathbf{s},\mathbf{w},\mathbf{y})$ and $p(\mathbf{y}|\mathbf{w},\mathbf{z},\mathbf{s})$. This section presents approximations to these distributions.

4.1 The Posterior of Latent Confounders

Exact inference of $\mathbf{z}$ is intractable for many models, such as multi-layered neural networks. Hence, we infer $\mathbf{z}$ using variational inference, which approximates the true posterior $p(\mathbf{z}|\mathbf{x},\mathbf{s},\mathbf{w},\mathbf{y})$ by a variational posterior $q(\mathbf{z}|\mathbf{x},\mathbf{s},\mathbf{w},\mathbf{y})$. This approximation is obtained by minimizing the Kullback-Leibler divergence (D_{KL}): $D_{\text{KL}}[q(\mathbf{z}|\mathbf{x},\mathbf{s},\mathbf{w},\mathbf{y})\|p(\mathbf{z}|\mathbf{x},\mathbf{s},\mathbf{w},\mathbf{y})]$, which is equivalent to maximizing the evidence lower bound (ELBO) of the marginal likelihood:

$$\mathcal{L} = E_{\mathbf{z}}\Big[\log p(\mathbf{y},\mathbf{w},\mathbf{s},\mathbf{x}|\mathbf{z})\Big] \qquad (7)$$
$$-\sum_{t=1}^{T} E_{\mathbf{z}_{t-1}}\Big[D_{\text{KL}}\big(q(\mathbf{z}_t|y_t,\mathbf{x}_t,w_t,\mathbf{s}_t)\|p(\mathbf{z}_t|\mathbf{z}_{t-1})\big)\Big].$$

The expectations are taken with respect to variational posterior $q(\mathbf{z}|\mathbf{x},\mathbf{s},\mathbf{w},\mathbf{y})$, and each term in the ELBO depends on the assumption of their distribution family presented in Section 3. We further assume that the variational posterior distribution takes the form $q(\mathbf{z}_t|\cdot) = \mathsf{N}(\mathbf{z}_t|f_q(y_t,w_t,\mathbf{s}_t,\mathbf{x}_t),\sigma_q^2\mathbf{I}_{d_z})$, where $f_q\colon \mathcal{Y}\times\mathcal{W}\times\mathcal{S}\times\mathcal{X}\to\mathcal{F}_z$ is a function parameterizing the designated distribution, σ_q is a hyperparameter, and $\mathbf{I}_{d_x}$ denotes the identity matrix. Our aim is to learn f_i where $i\in\boldsymbol{\mathcal{A}}\leftarrow\boldsymbol{\mathcal{A}}\cup\{q\}$.

4.1.1 Inference of The Posterior

To formulate a regularized empirical risk, we draw L samples of $\mathbf{z}$ from the variational posterior using reparameterization trick: $\mathbf{z}^l = [\mathbf{z}_1^l,...,\mathbf{z}_T^l]$ with $\mathbf{z}_t^l = f_q(y_t,w_t,\mathbf{s}_t,\mathbf{x}_t)+\sigma_q\boldsymbol{\varepsilon}_t^l$ and $\boldsymbol{\varepsilon}_t^l\sim\mathsf{N}(0,\mathbf{I}_{d_x})$. By drawing L noise samples $\boldsymbol{\varepsilon}_t^1,...,\boldsymbol{\varepsilon}_t^L$ at each time interval t in advance, we obtain a *complete* dataset

$$\mathcal{D} = \bigcup_{l=1}^{L}\bigcup_{t=1}^{T}\Big\{\,\big(y_t,w_t,\mathbf{x}_t,\mathbf{s}_t,\mathbf{z}_t^l\big)\,\Big\}.$$

At each time interval t, the dataset gives a tuple of the observed values $y_t,w_t,\mathbf{x}_t,\mathbf{s}_t$, and an expression of $\mathbf{z}_t^l = f_q(y_t,w_t,\mathbf{s}_t,\mathbf{x}_t)+\sigma_q\boldsymbol{\varepsilon}_t^l$. We state the following:

Lemma 1. *Let κ_i be kernels and $\mathcal{H}_i$ their associated reproducing kernel Hilbert space (RKHS), where $i\in\boldsymbol{\mathcal{A}}$. Let the empirical risk obtained from the negative ELBO be $\widehat{\mathcal{L}}$. Consider minimizing the following objective function*

$$J = \widehat{\mathcal{L}}\left(\bigcup\nolimits_{i\in\boldsymbol{\mathcal{A}}} f_i\right) + \sum_{i\in\boldsymbol{\mathcal{A}}}\lambda_i\|f_i\|_{\mathcal{H}_i}^2 \qquad (8)$$

with respect to functions f_i ($i\in\boldsymbol{\mathcal{A}}$), where $\lambda_i\in\mathbb{R}^+$. Then, the minimizer of (8) has the following form $f_i = \sum_{l=1}^{T\times L}\kappa_i(\,\cdot\,,\boldsymbol{\nu}_l^i)\boldsymbol{\beta}_l^i$ ($\forall i\in\boldsymbol{\mathcal{A}}$), where $\boldsymbol{\nu}_l^i$ is the l^{th} input to function f_i, i.e., it is a subset of the l^{th} tuple of $\mathcal{D}$, and the coefficients $\boldsymbol{\beta}_l^i$ are vectors in the Hilbert space $\mathcal{F}_i$. This minimizer further emits the following solution: $\boldsymbol{\beta}^i = [\boldsymbol{\beta}_l^i,...,\boldsymbol{\beta}_{TL}^i]^\top = \big(\sum_{l=1}^{L}\mathbf{K}_i^{l\top}\mathbf{K}_i^l + \lambda_i L\mathbf{K}_i\big)^{-1}\sum_{l=1}^{L}\mathbf{K}_i^{l\top}\psi^i$ for $i\in\boldsymbol{\mathcal{A}}\setminus\{w,q\}$ and $\boldsymbol{\psi}^y = \mathbf{y}$, $\boldsymbol{\psi}^x = \mathbf{x}$, $\boldsymbol{\psi}^s = \mathbf{s}$, $\boldsymbol{\psi}^z = \mathbf{K}_q\boldsymbol{\beta}^q$.

As mentioned in Section 3, we propose to model f_i using kernel methods. A natural way is to develop a slight modification to the classical representer theorem (Kimeldorf and Wahba, 1970; Schölkopf et al., 2001) so that it can be applied to the optimization of the ELBO. Eq. (8) is minimized with respect to the weights $\boldsymbol{\beta}^i$ and hyperparameters of the kernels. The proof of Lemma 1 is provided in Appendix.

Lemma 2. *For any fixed $\boldsymbol{\beta}^q$, the objective function J in Eq. (8) is convex with respect to $\boldsymbol{\beta}^i$ for all $i\in\boldsymbol{\mathcal{A}}\setminus\{q\}$.*

Proof. We present the sketch of proof here and details are deferred to Appendix. It can be shown that the objective function is a combination of several components including $(\boldsymbol{\beta}^i)^\top\mathbf{C}\boldsymbol{\beta}^i$, $\mathbf{c}^\top\boldsymbol{\beta}^i$, $-\mathbf{w}^\top\log\varphi(\mathbf{K}_w^l\boldsymbol{\beta}^w)$ and $-(\mathbf{1}-\mathbf{w})^\top\log\varphi(-\mathbf{K}_w^l\boldsymbol{\beta}^w)$, where $i\in\{y,s,x,z\}$, $\mathbf{C}$ is a positive semi-definite matrix and $\mathbf{c}$ is a vector. The first component is a quadratic form. Thus, its second-order derivative with respective to $\boldsymbol{\beta}^i$ is a positive semi-definite matrix; hence, it is convex. The second term is a linear function of $\boldsymbol{\beta}^i$ thus it is convex. The two last terms come from cross-entropy loss function and the input to these function are linear combination of the kernel function evaluated between each pair of data points. So these two terms are also convex. Consequently, J is convex with respective to $\boldsymbol{\beta}^i$ ($i\in\boldsymbol{\mathcal{A}}\setminus\{q\}$) because it is a linear combination of convex components. □

Lemma 2 implies that at an iteration in the optimization that $\boldsymbol{\beta}^q$ reaches its convex hull, the objective function J will reach its minimal point after a few more iterations. This is because the non-convexity of J is induced only by $\boldsymbol{\beta}^q$. This result indicates that we should attempt different random initialization on $\boldsymbol{\beta}^q$ instead of the other parameters when optimizing J because $\boldsymbol{\beta}^i$ ($i\in\boldsymbol{\mathcal{A}}\setminus\{q\}$) always converges to its optimal point (conditioned on $\boldsymbol{\beta}^q$).

4.2 The Auxiliary Distributions

The previous steps approximate the posterior $p(\mathbf{z}|\cdot)$ by variational posterior $q(\mathbf{z}|\cdot)$ and estimate the density of $p(\mathbf{y}|\mathbf{w}, \mathbf{z}, \mathbf{s})$. This section outlines the approximation of $p(\mathbf{y}|\mathbf{s}, \mathbf{x}, \mathbf{w})$, $p(\mathbf{w}|\mathbf{s}, \mathbf{x})$ and $p(\mathbf{s}|\mathbf{x})$. Denoting their corresponding approximation as $\widetilde{p}(\mathbf{y}|\mathbf{s}, \mathbf{x}, \mathbf{w})$, $\widetilde{p}(\mathbf{w}|\mathbf{s}, \mathbf{x})$, $\widetilde{p}(\mathbf{s}|\mathbf{x})$, we estimate the parameters of those distributions directly using classical representer theorem.

In the following, we briefly describe how to approximate $\widetilde{p}(\mathbf{w}|\mathbf{s}, \mathbf{x})$. The regularized empirical risk obtained from the negative log-likelihood of $\widetilde{p}(\mathbf{w}|\mathbf{s}, \mathbf{x})$ is as follows:

$$J_w = \sum_{t=1}^{T} \ell_{\text{Xent}}\big(w_t, g_w(w_{t-1}, \mathbf{s}_t, \mathbf{x}_t)\big) + \delta_w \|g_w\|^2_{\mathcal{V}_w},$$

where $\ell_{\text{Xent}}(\cdot, \cdot)$ is the cross-entropy loss function, $\delta_w \in \mathbb{R}^+$, $g_w \colon \mathcal{W} \times \mathcal{S} \times \mathcal{X} \mapsto \mathcal{F}_w$, and $\mathcal{V}_w$ is the RKHS with kernel function $\tau_w(\cdot, \cdot)$. By classical representer theorem, the minimized form of g_w is given by:

$$g_w = \sum_{j=1}^{T} \alpha_j^w \tau_w(\,\cdot\,, [w_{j-1}, \mathbf{s}_j, \mathbf{x}_j]),$$

where $\alpha_j^w \in \mathbb{R}$ is the parameter to be learned. It can be shown that J_w is a convex objective function because the input to the cross-entropy function is linear. Other distributions can be estimated in a similar fashion and the full description is provided in Appendix.

5 EXPERIMENTS

Baselines and aims of experiments. In this section, we examine the performance of the proposed CausalSC in estimating causal effects from temporal data on both synthetic and real-world datasets. We compare with the following baselines: **(1)** The potential outcomes-based model for time series data by Bojinov and Shephard (2019), which uses Horvitz-Thompson estimator to evaluate causal effects. The key factor of this method is the 'adapted propensity score'; we have implemented two versions of this score. The first one uses a fully connected neural network. Herein, we assume that $p(w_t|\mathbf{w}_{1:t-1}, \mathbf{y}_{1:t-1}) = \mathsf{Bern}(w_t|f(w_{t-1}, y_{t-1}))$ with $f(w_{t-1}, y_{t-1})$ is a neural network taking the observed values of w_{t-1}, y_{t-1} as input to predict w_t. The second one uses Long-Short Term Memory (LSTM) to estimate $p(w_t|\mathbf{w}_{1:t-1}, \mathbf{y}_{1:t-1})$. **(2)** The second baseline is CFRNet, a model for inferring treatment effects by Shalit et al. (2017). **(3)** The third baseline, CEVAE (Louizos et al., 2017) is a causal inference model based on variational auto-encoders. We use the code of CEVAE and CFRNet which are available online to train these models. **(4)** The last baseline is fairness through causal awareness (FCA) by Madras et al. (2019). This method is an extension of CEVAE that considers two types of confounder: observed and latent ones.

For the neural network setup on each baseline, we closely follow the architecture of Louizos et al. (2017) and Shalit et al. (2017). Unless otherwise stated, we utilize a fully connected network with the exponential linear unit (`ELU`) activation function (Clevert et al., 2016) and use the same number of nodes in each hidden layer. We fine-tune the network with $2, 4, 6$ hidden layers and $50, 100, 150, 200, 250$ nodes per layer. We also fine-tune the learning rate in $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ and use *Adamax* (Kingma and Ba, 2015) for optimization.

Evaluation metrics. To evaluate the performance of each method, we report the absolute error of the ATE: $\epsilon_{\text{ATE}} := |\,\text{ATE} - \widehat{\text{ATE}}\,|$ and the precision of estimating heterogeneous effects (PEHE) (Hill, 2011): $\epsilon_{\text{PEHE}} := \big(\sum_{i=T_1}^{T_2} (\text{EP}_i - \widehat{\text{EP}}_i)^2\big)/(T_2 - T_1 + 1)$.

5.1 Illustration on Modeling with Stochastic Confounders

Before carrying out our main experiments, we first illustrate the importance of our proposed method in estimating causal effects for time series data with stochastic confounders. We consider the ground truth causal model whose structural equations are as follows:

$$\begin{aligned} s_t &= a_0 + a_1 s_{t-1} + o_t, \\ w_t &= \mathbb{1}(\varphi(b_0 + b_1 z_t) \geq u_t), \\ y_t &= f_y(s_t, w_t) + v_t, \end{aligned}$$

where $\mathbb{1}(\cdot)$ is the indicator function, $s_0 = 0$, $o_t \sim \mathsf{N}(0, 0.3^2)$, $u_t \sim \mathsf{U}[0, 1]$, and $v_t \sim \mathsf{N}(0, 1)$. In this model, s_t, w_t, y_t are endogenous variables, and o_t, u_t, v_t are exogenous variables. The functions f_i $(i \in \boldsymbol{\mathcal{A}} \setminus \{y\})$ in this model are linear. We consider three cases for the ground truth of function f_y:

(1). Linear: $f_y(s_t, w_t) = c_0 + c_1 s_t + c_2 w_t$.
(2). Quadratic: $f_y(s_t, w_t) = (c_0 + c_1 s_t + c_2 w_t)^2$.
(3). Exponential: $f_y(s_t, w_t) = \exp(c_0 + c_1 s_t + c_2 w_t)$.

For all the cases, we randomly choose the true parameters of the models as follows $(a_0, a_1, b_0, b_1, c_0, c_1, c_2) = (0.70, 0.95, 0.20, -0.10, 0.70, 0.40, 1.70)$, and then sample three time series of length $T = 1000$ for s_t, w_t, y_t from the above distributions. Herein, we keep w_t, y_t, s_t as the observed data and use the following three inference models to estimate causal effects: Model-1: without confounders, Model-2: with *iid* confounders and Model-3: with stochastic confounders. The errors reported in Table 1 show that the model with stochastic confounders outperforms the others as it fits data well. In the following sections, we present our

Table 1: The errors of the estimated treatment effects (lower is better).

	Model-1 without confounders		Model-2 with *iid* confounders		Model-3 CausalSC	
	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}
Linear outcome	0.06±.01	0.06±.01	**0.05±.01**	**0.05±.01**	**0.05±.01**	**0.05±.01**
Quadratic outcome	2.53±.08	2.26±.06	1.78±.03	0.88±.03	**0.33±.02**	**0.26±.02**
Exponential outcome	4.52±.12	3.36±.15	4.18±.20	2.41±.07	**0.91±.03**	**0.78±.06**

main experiments with more complicated functions f_i ($i \in \boldsymbol{\mathcal{A}}$).

5.2 Synthetic Experiments

Since obtaining the ground truth for evaluating causal inference methods is challenging, most of the recent methods are evaluated with synthetic or semi-synthetic datasets (Louizos et al., 2017). This set of experiments is conducted on four synthetic datasets and one benchmark dataset, where three synthetic datasets are time series with latent stochastic confounders and the other two datasets are static data with *iid* confounders.

Synthetic dataset. We simulate data for y_t, w_t, $\mathbf{x}_t$, $\mathbf{s}_t$ and $\mathbf{z}_t$ from their corresponding distributions as in Eqs. (1)-(3), each with length $T = 200$. The ground truth nonlinear functions $f_i (i \in \boldsymbol{\mathcal{A}} \setminus \{q\})$ with respect to the distributions of $y_t, w_t, \mathbf{x}_t, \mathbf{s}_t, \mathbf{z}_t$ are fully connected neural networks (refer to Appendix for details of these functions). Using different numbers of the hidden layers, i.e., 2, 4, and 6, we construct three synthetic datasets, namely TD2L, TD4L, and TD6L. For these three datasets, we sample the latent confounder variable $\mathbf{z}$ satisfying Eq. (1). We also construct another dataset, TD6L-iid, that uses 6 hidden layers but with the *iid* latent confounder variable $\mathbf{z}$, i.e., $\mathbf{z}_t \perp\!\!\!\perp \mathbf{z}_s, \forall t, s$. Finally, we only keep y_t, w_t, $\mathbf{x}_t$, $\mathbf{s}_t$ as observed data for training. Further details of the simulation are provided in Appendix.

Benchmark dataset. Infant Health and Development Program (IHDP) dataset (Hill, 2011) is a study on the impact of specialist visits on the cognitive development of children. This dataset has 747 entries, each with 25 covariates. The treatment group consists of children who received specialist visits and a control group that includes children who did not. For each child, a treated and a control outcome were simulated from the numerical scheme of the NPCI library Dorie (2016).

Results and discussion. Each of the experimental datasets has 10 replications. For each replication, we use the first 64% for training, the next 20% for validation, and the last 16% for testing. We examine three setups for our inference method, one with kernel method to model the nonlinear functions, another one with the neural networks, and the third one is an iid confounder setting with kernel method to model the nonlinear functions. Specifically, 'CausalSC-Matérn kernel', 'CausalSC-RBF kernel' and 'CausalSC-RQ kernel' denote our proposed method with representer theorem to model the nonlinear functions f_i ($i \in \boldsymbol{\mathcal{A}}$) using Matérn kernel, radial basis function kernel, and rational quadratic function kernel, respectively. CausalSC-NNjL denotes our method using neural networks to model these nonlinear functions, where $j \in \{2, 4, 6\}$ is the number of hidden layers. 'iid confounder-Matérn kernel', 'iid confounder-RBF kernel' and 'iid confounder-RQ kernel' denote our proposed method with representer theorem and independent confounders.

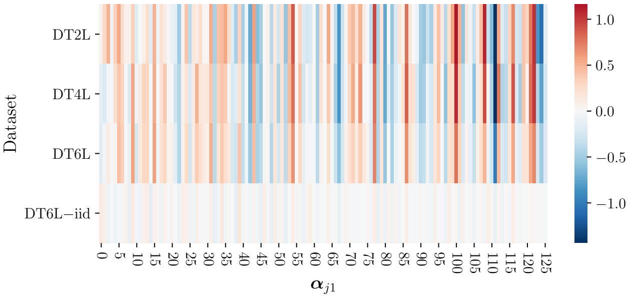

Figure 3: The heatmap of $\boldsymbol{\alpha}_{j1}$ on each dataset.

Table 2 reports the errors of each method, where significant results are highlighted in bold. We observe that the performance of our model is competitive for the first three datasets since our framework is *suited* and built for temporal data. This verifies the effectiveness of our proposed method on the inference of the causal effects for temporal data, especially with latent stochastic confounders. Moreover, the use of representer theorem returns *similar values* of ATE on three different kernel functions. Our proposed method outperforms the other baselines on the first three datasets, this is because we consider the time-dependency in the latent confounders, while the others do not take into account such property. For datasets that respects *iid* confounder, our methods give comparable results with the other baselines (the last five lines in Table 2). This can be explained as follows. In our setup, $\mathbf{z}_t$ has the following form:

Table 2: Out-of-sample ATE error (ϵ_{ATE}) of each method on different datasets (lower is better).

Method	Temporal Data (latent stochastic confounders)			IID Data (*iid* confounders)	
	TD2L	TD4L	TD6L	TD6L-iid	IHDP
CausalSC-Matérn kernel	**0.299±.069**	**0.193±.088**	**0.237±.056**	0.412±.116	**0.290±.076**
CausalSC-RBF kernel	**0.341±.078**	**0.289±.025**	**0.287±.067**	0.397±.096	0.460±.130
CausalSC-RQ kernel	**0.311±.067**	**0.255±.020**	**0.257±.063**	**0.342±.076**	0.489±.170
CausalSC-NN2L	**0.369±.080**	0.395±.162	0.396±.169	0.390±.107	4.057±.109
CausalSC-NN4L	0.413±.092	**0.309±.063**	0.313±.060	0.398±.104	1.048±.441
CausalSC-NN6L	0.477±.102	0.325±.069	**0.297±.070**	**0.388±.106**	0.306±.063
iid confounder-Matérn kernel	0.519±.072	0.658±.098	0.762±.068	**0.311±.092**	**0.217±.095**
iid confounder-RBF kernel	0.632±.085	0.982±.120	0.884±.078	**0.346±.075**	**0.277±.091**
iid confounder-RQ kernel	0.640±.085	1.081±.103	0.785±.077	**0.335±.076**	0.283±.075
Bojinov and Shephard (2019) (FC)	1.477±.220	1.243±.219	1.180±.177	0.470±.081	0.529±.183
Bojinov and Shephard (2019) (LSTM)	1.316±.261	1.117±.232	1.179±.323	0.593±.102	0.613±.137
Shalit et al. (2017) (CFRNet)	0.780±.131	0.570±.098	0.701±.131	0.568±.091	0.424±.100
Louizos et al. (2017) (CEVAE)	1.166±.247	1.428±.709	0.705±.179	**0.337±.093**	**0.232±.061**
Madras et al. (2019) (FCA)	0.391±.078	1.065±.504	0.682±.080	0.393±.103	**0.261±.063**

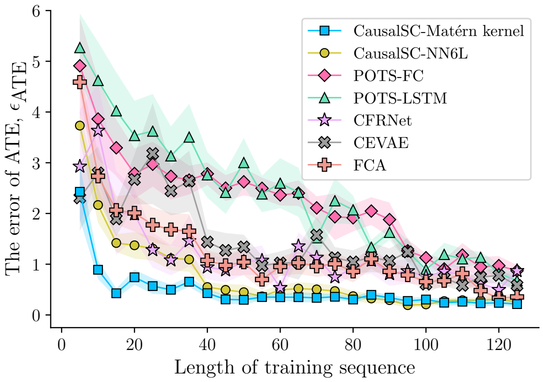

Figure 4: ϵ_{ATE} on different lengths of the training set (lower is better).

$\mathbf{z}_t = \boldsymbol{\alpha}_0 + \sum_{j=1}^{T}\sum_{l=1}^{L} \boldsymbol{\alpha}_{jl}\, k_z(\mathbf{z}_{t-1}, \mathbf{z}_j^l) + \boldsymbol{\epsilon}_t$, where $\boldsymbol{\alpha}_0$ is a bias vector, $\boldsymbol{\alpha}_{jl}$ is a weight vector, and $\boldsymbol{\epsilon}_t$ is the Gaussian noise (Please refer to Appendix for specific expressions of $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_{jl}$). The learned weights $\boldsymbol{\alpha}_{jl}$ presented in Figure 3 (for $l = 1$) show that their quantities are around 0 on data with *iid* confounders, which breaks the connection from $\mathbf{z}_{t-1}$ to $\mathbf{z}_t$ and thus makes these two variables independent to each other.

Figure 4 presents the convergence of each method on the dataset TD6L, over different lengths of the training set $T_{\text{train}} \in \{5, 10, ..., 125\}$. In general, the more training data we have, the smaller the error of the estimated ATE. The figure reveals that our method (blue line) starts to converge from around $T_{\text{train}} = 45$, which is faster than the others. Additionally, the estimated ATE of our method is stable with a small error bar.

5.3 Real Data: Gold–Oil Dataset

Data description. Gold and oil are among the most transactable commodities in real-life. An increasing trend of oil price would result in a rise of gold price since it may generate higher inflation that pushes up the demand for gold (Le and Chang, 2011; Šimáková, 2011). In this section, we examine the causal effects from the price of crude oil to that of gold. The dataset in this experiment consists of monthly prices of some commodities including gold, crude oil, beef, chicken, cocoa-beans, rice, shrimp, silver, sugar, gasoline, heating oil and natural-gas from May 1989 to May 2019. We consider the price of gold as the outcome $\mathbf{y}$, and the trend of crude oil's price as the treatment $\mathbf{w}$. Specifically, we cast an increase of crude oil's price to 1 ($w_t = 1$) and a decrease of crude oil's price to 0 ($w_t = 0$). We use the prices of gasoline, heating oil, natural-gas, beef, chicken, cocoa-beans, rice, shrimp, silver and sugar as proxy variables $\mathbf{x}$.

To make it suitable for our setting, the following preprocessing procedure is performed. Let $\nu_0, \nu_1, \nu_2, ..., \nu_T$ be time series of a commodity. We take the difference between two consecutive observations as input to our model, i.e., $\Delta\nu_t = \nu_t - \nu_{t-1}$. This preprocessing is applied to all the prices of commodities, thus the inputs to our model are actually the differenced series: $\Delta\nu_1, \Delta\nu_2, \Delta\nu_3, ..., \Delta\nu_T$. For the treatment w_t ($t = 1, 2, ..., T$), we further cast it to a binary value, i.e., $w_t = \mathbb{1}(\Delta\nu_t > 0)$, where $\mathbb{1}(\cdot)$ denotes the indicator function. By using the differenced series, we are removing trends and seasonal effects (see, e.g, Commandeur and Koopman, 2007, Chapter 10; Jinka and Schwartz, 2016, Chaper 4). Hence, it would be more reasonable to assume that the preprocessed data satisfies our proposed causal graph, i.e., the two consecutive differenced values of the observations are independent given the latent confounders, $\Delta\nu_t \perp\!\!\!\perp \Delta\nu_{t-1} \,|\, \mathbf{z}_t$.

Results and discussion. We evaluate the effect path and ATE between two sequences of treatments

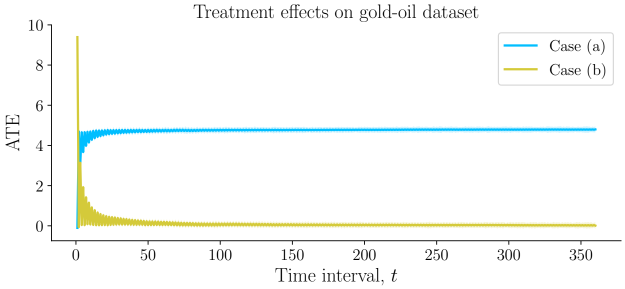

Figure 5: *Case* (a): ATE between two treatment paths $[1, 1, ..., 1, 1]^\top$ and $[0, 1, ..., 0, 1]^\top$, *Case* (b): ATE between two treatment paths $[1, 0, ..., 1, 0]^\top$ and $[0, 1, ..., 0, 1]^\top$.

$\mathbf{w}_1 = [1, 1, ..., 1, 1]^\top$ (increasing crude oil prices) and $\mathbf{w}_2 = [0, 1, ..., 0, 1]^\top$ (alternating decreasing and increasing crude oil prices, i.e., constant on average). In Figure 5, Case (a) presents the estimated ATE of the above two sequences of treatments. The estimated ATE using CausalSC is 4.8, which means that in a period the price of crude oil increases, the average gold price is about to increase 4.8. This quantity is equivalent to an increase of 0.77% in the gold price over the period $(4.8/\{(\sum_{t=1}^{T} y_t^{obs})/T\} = 0.77\%)$. To validate the 0.77% increase in gold price, we compare this with the results reported in Šimáková (2011) (based on Granger causality) that show the "percentage increase in oil price leads to a 0.64% increase in gold price". We note that our results give similar order of magnitude and the slight difference may be attributed to our experimental data that is from May 1989 to May 2019, while the analysis in Šimáková (2011) is on the data from 1970 to 2010.

In Figure 5, Case (b), we present another experiment with $\mathbf{w}_1 = [1, 0, ..., 1, 0]^\top$ and $\mathbf{w}_2 = [0, 1, ..., 0, 1]^\top$. In this case, both treatment paths represent the alternating variation of crude oil prices. Specifically, the former increases first and then decreases, while the latter is on the opposite. The average treatment effect is expected to be around 0. From Figure 5, Case (b), the estimated ATE by CausalSC is 0.0045, which is in line with the expectation. To check on statistical significance, we performed a one group t-test on the EP (Definition 1) with the population mean to be tested is 0. The p-value given by the t-test is 0.9931, which overwhelmingly fail to reject the null hypothesis that the ATE equals 0. This again verifies the reasonable performance of the proposed method.

6 CONCLUSION

We have developed a causal modeling framework, CausalSC, that admits observed and latent confounders as random processes, generalizing recent work where the confounders are assumed to be independent and identically distributed. We study the causal effects over time using variational inference in conjunction with an alternative form of the representer theorem with a random input space. Our algorithm supports causal inference from the observed outcomes, treatments, and covariates, without parametric specification of the components and their relations. This property is important for capturing real-life causal effects in SCM, where non-linear functions are typically placed in the priors. Our setup admits non-linear functions modulating the SCM with estimated parameters that have analytical solutions. This approach compares favorably to recent techniques that model similar non-linear functions to estimate the causal effects with neural networks, which usually involve extensive model tuning and architecture building.

One limitation of our framework is that the fixed amount of passing time (time-lag) is set to unity as it leads to further simplifications in computing causal effects. Of practical interest is to perform a more detailed empirical study on general time-lag.

Acknowledgements

We wish to thank Young Lee (Harvard University) for discussions related to this work. This work was partially supported by a Research Scholarship and an Academic Research Grant No. T1 251RES1827 from the Ministry of Education, Singapore.

References

Alaa, A. M. and van der Schaar, M. (2017). Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, pages 3424–3432.

Bahadori, M. T. and Liu, Y. (2012). On causality inference in time series. In *2012 AAAI Fall Symposium Series*.

Bica, I., Alaa, A. M., Jordon, J., and van der Schaar, M. (2020a). Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations*.

Bica, I., Alaa, A. M., and van der Schaar, M. (2020b). Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. In *Proceedings of the 37th International Coference on International Conference on Machine Learning*.

Bojinov, I. and Shephard, N. (2019). Time series experiments and causal estimands: exact randomization tests and trading. *Journal of the American Statistical Association*, pages 1–36.

Clevert, D.-A., Unterthiner, T., and Hochreiter, S. (2016). Fast and accurate deep network learning by exponential linear units (elus). In *Proceedings of the 4th International Conference on Learning Representations.*

Commandeur, J. J. and Koopman, S. J. (2007). *An introduction to state space time series analysis.* Oxford University Press.

de Luna, X., Fowler, P., and Johansson, P. (2017). Proxy variables and nonparametric identification of causal effects. *Economics Letters*, 150:152–154.

Dorie, V. (2016). Npci: Non-parametrics for causal inference. *URL: https://github. com/vdorie/npci.*

Duncan, O. D. (1975). *Introduction to Structural Equation Models.* New York: Academic Press.

Duncan, O. D. (2014). *Introduction to structural equation models.* Elsevier.

Eichler, M. (2005). A graphical approach for evaluating effective connectivity in neural systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1457):953–967.

Eichler, M. (2007). Granger causality and path diagrams for multivariate time series. *Journal of Econometrics*, 137(2):334–353.

Eichler, M. (2009). Causal inference from multivariate time series: What can be learned from granger causality. In *13th International Congress on Logic, Methodology and Philosophy of Science.*

Eichler, M. and Didelez, V. (2010). On granger causality and the effect of interventions in time series. *Lifetime data analysis*, 16(1):3–32.

Frees, E. W. et al. (2004). *Longitudinal and panel data: analysis and applications in the social sciences.* Cambridge University Press.

Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica: Journal of the Econometric Society*, pages 979–1001.

Goldberger, A. S. (1973). Structural equation models: An overview. *Structural equation models in the social sciences*, pages 1–18.

Granger, C. W. (1980). Testing for causality: a personal viewpoint. *J. Economic Dynamics and control*, 2:329–352.

Guo, R., Cheng, L., Li, J., Hahn, P. R., and Liu, H. (2020). A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, 53(4):1–37.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.

Huang, B., Zhang, K., Gong, M., and Glymour, C. (2019). Causal discovery and forecasting in nonstationary environments with state-space models. In *Proceedings of the 36th International Conference on Machine Learning.*

Jabbari, F., Ramsey, J., Spirtes, P., and Cooper, G. (2017). Discovery of causal models that contain latent variables through bayesian scoring of independence constraints. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 142–157. Springer.

Jinka, P. and Schwartz, B. (2016). *Anomaly Detection for Monitoring.* O'Reilly Media, Incorporated.

Kamiński, M., Ding, M., Truccolo, W. A., and Bressler, S. L. (2001). Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biological cybernetics*, 85(2):145–157.

Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.*

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165.

Kuroki, M. and Pearl, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437.

Le, T.-H. and Chang, Y. (2011). Oil and gold: correlation or causation? Technical report, Nanyang Technological University, School of Social Sciences, Economic Growth Centre.

Li, S. and Bühlmann, P. (2018). Estimating heterogeneous treatment effects in nonstationary time series with state-space models. *arXiv preprint arXiv:1812.04063.*

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. (2017). Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456.

Lu, C., Schölkopf, B., and Hernández-Lobato, J. M. (2018). Deconfounding reinforcement learning in observational settings. *arXiv preprint arXiv:1812.10576*.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. (2019). Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 349–358. ACM.

Montgomery, M. R., Gragnolati, M., Burke, K. A., and Paredes, E. (2000). Measuring living standards with proxy variables. *Demography*, 37(2):155–174.

Neyman, J. S. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9. (translated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480). *Annals of Agricultural Sciences*, 10:1–51.

Nie, X. and Wager, S. (2020). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*.

Ning, B., Ghosal, S., Thomas, J., et al. (2019). Bayesian method for causal inference in spatially-correlated multivariate time series. *Bayesian Analysis*, 14(1):1–28.

Oprescu, M., Syrgkanis, V., and Wu, Z. S. (2019). Orthogonal random forest for causal inference. In *International Conference on Machine Learning*, pages 4932–4941. PMLR.

Pearl, J. (1988). Probabilistic reasoning in intelligent systems. *San Mateo, CA: Kaufmann*, 23:33–34.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Pearl, J. (2000). *Causality: models, reasoning and inference*, volume 29. Springer.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.

Peters, J., Janzing, D., and Schölkopf, B. (2013). Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems*, pages 154–162.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053.

Riegg, S. K. (2008). Causal inference and omitted variable bias in financial aid research: Assessing solutions. *The Review of Higher Education*, 31(3):329–354.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer.

Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3076–3085.

Šimáková, J. (2011). Analysis of the relationship between oil and gold prices. *Journal of finance*, 51(1):651–662.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. (2000). *Causation, prediction, and search*. MIT press.

Tian, J. and Pearl, J. (2001). Causal discovery from changes. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, page 512–521, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. (2018). Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pages 2633–2643.

Yoon, J., Jordon, J., and van der Schaar, M. (2018). GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*.