Time Series Causal Link Estimation under Hidden Confounding using Knockoff Interventions

Anonymous Author(s)
Affiliation
Address
email

Abstract

Latent variables often mask cause-effect relationships in observational data which provokes spurious links that may be misinterpreted as causal. This problem sparks great interest in the fields such as climate science and economics. We propose to estimate confounded causal links of time series using Sequential Causal Effect Variational Autoencoder (SCEVAE) while applying knockoff interventions. We show the advantage of knockoff interventions by applying SCEVAE to synthetic datasets with both linear and nonlinear causal links. Moreover, we apply SCEVAE with knockoffs to real aerosol-cloud-climate observational time series data. We compare our results on synthetic data to those of a time series deconfounding method both with and without estimated confounders. We show that our method outperforms this benchmark by comparing both methods to the ground truth. For the real data analysis, we rely on expert knowledge of causal links and demonstrate how using suitable proxy variables improves the causal link estimation in the presence of hidden confounders.

1 Introduction

Causal link estimation and inference for dynamical systems are important tasks in many fields such as finance and climate science (1). We propose to estimate both linear and nonlinear causal effects of time series variables under hidden confounding using Sequential Causal Effect Variational Autoencoder (SCEVAE) (2) while performing knockoff interventions. SCEVAE is based on a Causal Effect Variational Autoencoder (CEVAE) by Louizos et al. (3), a deep learning framework which was first applied to analyzing effects of a binary treatment on the patients' health outcomes. The causal Directed Acyclic Graph (DAG) representing hidden confounding with one proxy variable, as used by CEVAE framework, is depicted in Fig. 1. Y denotes the effect, W the cause, Z denotes the hidden confounder and X a proxy variable providing noisy views on Z. We note that the proxy can be multivariate, and both categorical and continuous. Although the DAG in Fig. 1 might appear restrictive, in many real-world applications, such as climate science for instance, there are unobserved confounders present, with multiple different proxy variables to describe them (4).

We assume that the causal link between the confounded variables is approximately directly identifiable through the proxy to the hidden confounder. We provide more details on this and the identifiability theory in Section 3.2.

Knockoff filters were first introduced by Barber and Candès in 2015 (5) for controlling the variable selection procedure and false discovery rate of potential drivers of the response variable. Their method produces knockoff variables which imitate the correlation structure of the existing variables and do not require any new data. At first they were limited to a linear Gaussian response variable, but this was later extended for arbitrary and unobserved data distributions in the work by Romano et al. (6), as they introduced a framework called Deep Knockoffs based on deep generative models.

Submitted to *A causal view on dynamical systems* workshop at NeurIPS 2022. Do not distribute.

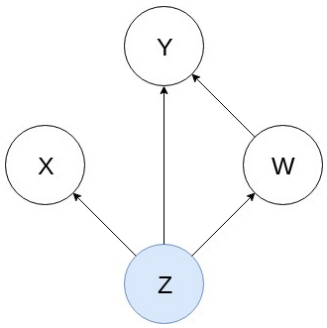

Figure 1: Causal DAG with observed variables $\{W, X, Y\}$ and latent variable $\{Z\}$.

In the work by (2), the authors used Gaussian noise as an intervention since it removes the link to the hidden confounder as required by do-calculus (7), but it does not preserve the distribution of the original cause variable W. We therefore propose to use a knockoff (5) of W as the intervention. In addition to preserving the distribution of W, this also preserves the inter-variable correlations to other observed variables, but is not correlated to W. One limitation of our approach is that we assume that, within the processed window, data is stationary and sufficient to produce good knockoffs.

Apart from the nonlinear causal links, the problem of causal link estimation under hidden confounding is particularly challenging for time series because the data is non-stationary, and the causal link between confounded variables might be changing over time, thus introducing delayed causal effects.

The paper is structured as follows. In Section 2, we outline related work on treatment effect estimation and provide an overview of causality analysis under hidden confounding. Moreover, we discuss other causality analysis work that utilizes knockoffs. In Section 3, we introduce the CEVAE framework and provide definition and discussion of causal identifiability, the assumptions under which using SCEVAE is justified, as well as a detailed description of SCEVAE methodology with knockoff interventions and introduce knockoffs themselves in more detail. Section 4 describes the data we use, and 5 provides a detailed experimental setup along with our method's results. Section 6 summarizes our approach and concludes the paper.

To the best of our knowledge, we are the first to propose causal effect estimation under hidden confounding for sequential data using SCEVAE with knockoff interventions.

2 Related Work

Treatment effect estimation in the static setting is a very active field of research (8), (9), (10). Under the assumption of no hidden confounding, methods such as VCNet (11) and its extensions by state-of-the-art transformer networks (12) were recently applied to the task of continuous treatment estimation by Zhang et al. (13) and Melnychuk et al. (14).

In a more realistic scenario when there are unobserved confounders present, Causal Effect Variational Autoencoder (CEVAE) (3) was one of the first such deep learning methods. It relies on the existence of one or more proxies to directly estimate the latent confounder. Nevertheless, in contrast to above-mentioned transformer-based treatment estimation methods, CEVAE focuses on binary treatment. Research endeavors such as those by Rissanen and Marttinen (15), Im et al. (16), and Trifunov et al. (17) either extend CEVAE to a uniform or continuous treatment, or analyse this deep latent variable model's capabilities. However, Sequential CEVAE (SCEVAE) (2) is the first method to extend CEVAE to time series.

Taking a different approach towards treatment effect estimation, Bica et al. (18) and Hatt and Feuerriegel (19) propose deconfounding methods for sequential data using neural networks. Specifically, in the work by Bica et al. (18), the authors develop Time Series Deconfounder (TSdeconf), a method built upon RNNs with multitask output to produce a factor model over time and estimate latent variables. These latent variable estimates are then used for causal inference as proxies. The limitation of this approach, in contrast to SCEVAE, is that it requires many patients i.e. samples and is not suitable for processing long time series (e.g. $N \geq 1000$). Another shortcoming of TSdeconf is that

it cannot be applied to observational data with one or more treatments assigned at each time step. This was addressed by Hatt et al. (19), as the authors introduced a similar time series deconfounding method called Sequential Deconfounder, this time based on a Gaussian process latent variable model.

These methods are useful when there are no proxies available. In contrast to SCEVAE with knockoff intervention, this may introduce more approximation error and requires multiple realizations of each variable i.e. many independent samples for training. Although SCEVAE relies on access to proxies, it is not a limitation in many fields such as environmental and climate science, where many observed variables can be used to describe a latent one (4).

In the recent years, due to their distribution- and correlation-preserving properties, knockoffs (5), (6) were innovatively applied for counterfactual analysis of images (20), and time series (21), (22). In Ahmad et al. (21), the authors propose to use deep learning and counterfactual knockoff variables to aid causal discovery of multivariate nonlinear time series with lower false discovery rate than state-of-the-art methods.

Similarly to SCEVAE, work by Yin and Barucca (23) relies on RNNs to estimate the causal link of the variables under influence of a hidden confounder. However, it does not employ do-calculus (7), but rather Granger causality (24) to determine the presence of a non-spurious causal link. Due to this choice of causality analysis tools, it cannot detect instantaneous causal links. Moreover, since we use knockoff interventions the cause variable's distribution does not change, and its correlation to other observation variables is preserved. This type of intervention allows for more accurate causal link estimates in comparison to standard normal interventions originally proposed for SCEVAE approach as well as lower counterfactual prediction error as we will discuss in more detail in Section 5.1.

3 Methodology

3.1 Causal effect variational autoencoder

To better understand SCEVAE architecture, we first introduce Causal effect variational autoencoder (CEVAE) (3). It is a deep learning method based on a VAE (25) and a TARnet (26). Its underlying probabilistic graphical model is shown in Fig. 1. CEVAE methodology assumes all variables to be non-sequential. W denotes binary treatment, Y an outcome of this treatment, while the latent confounder Z, and its proxy X denote the socio-economic status and income of each patient, respectively. The central aim of treatment effect estimation is recovering the Individual Treatment Effect (ITE) and the Average Treatment Effect (ATE) defined in (1) and (2), respectively:

$$ITE(x) := \mathbb{E}_Y(Y|X = x, do(W = w^1)) - \mathbb{E}_Y(Y|X = x, do(W = w^0)) \tag{1}$$

$$ATE := \mathbb{E}_Y(ITE(x)) \tag{2}$$

These metrics are defined for each value x of variable X, and by w^1 we denote applied treatment, while values of W when no treatment is applied are denoted by w^0. In the CEVAE framework, $w^1 = 1$, and $w^0 = 0$. ATE is easily calculated once we obtain the ITE, and for that we need to recover the joint distribution $p(Z, X, W, Y)$, as stated in Theorem 1 by Louizos et al. (3).

Theorem 1. *If CEVAE recovers $p(Z, X, W, Y)$, then we can recover the ITE under the causal model in Fig. 1.*

Distribution $p(Z, X, W, Y)$ is obtained via CEVAE's model network by approximating the true posterior over Z conditioned on X, W and Y, whereas the prior $p(Z)$ is modeled by the standard normal distribution. All estimated probability distributions are parameterized by MLPs. TARnet is used to infer the estimate of the posterior by branching for each of the two treatment groups in W. Further, once all necessary distributions have been modeled, one can then construct a single objective for the inference and model networks, i.e. the *variational lower bound*:

$$\mathcal{L} = \sum_{i=1}^{N} \mathbb{E}_{q(z_i|x_i,w_i,y_i)}(\log p(z_i) - \log q(z_i|x_i, w_i, y_i) + \log p(x_i, w_i|z_i) + \log p(y_i|w_i, z_i)) \tag{3}$$

of the causal graphical model from Fig. 1. By x_i we denote an input data point, by w_i the treatment assignment, by y_i the outcome of the specific treatment, by z_i the latent confounder, and by $q(\cdot)$ we denote estimation of the probability distribution with the same arguments.

Finally, since it is necessary to know the intervention assignment W together with its outcome Y before inferring the posterior distribution over Z, two auxiliary distributions are introduced, helping to predict w_i and y_i for new samples, so the variational lower bound becomes

$$\mathcal{F}_{CEVAE} = \mathcal{L} + \sum_{i=1}^{N} (\log q(w_i = w_i^* | x_i^*) + \log q(y_i = y_i^* | x_i^*, w_i^*)), \tag{4}$$

where x_i^*, w_i^*, y_i^* are the observed values for the input, intervention, and outcome variables in the training set.

3.2 Causal identifiability

When estimating causal link intensity or performing causal discovery from observational data, one needs to establish if the underlying model is identifiable. If that is not the case, a set of assumptions under which the identifiability holds must be imposed. We introduce identifiability following recent work by Khemakhem et al. (27) and discuss this issue for deep latent variable models such as VAE.

Definition 1. Let $\sim$ be an equivalence relation on the set of model parameters Θ. We say that a deep latent variable model $p_\theta(Y, Z) = p_\theta(Y|Z)p_\theta(Z)$, for observed variable $Y \in \mathbb{R}^d$ and latent random vector $Z \in \mathbb{R}^n$ is *identifiable* up to $\sim$ if

$$p_\theta(Y) = p_{\hat{\theta}}(Y) \Rightarrow \theta \sim \hat{\theta} \tag{5}$$

for $\theta, \hat{\theta} \in \Theta$. The elements of the quotient space $\Theta/_\sim$ are called the *identifiability classes*.

The graphical model in Fig. 1 is in general not identifiable. A starting point would be to obtain model parameters or estimate latent variables $\mathbf{Z}^*$ up to trivial transformations T_i for $i \in \{1, \ldots, N\}$ in the form of sufficient statistics, and invertible matrix A (27).

To do that, we will first introduce sufficient statistics T_i with respect to the deep latent variable model from Definition 1 and illustrate the procedure on a one-dimensional latent variable case.

A statistic is sufficient for a family of probability distributions if the sample from which it is obtained provides no additional information than the statistic, as to which of those probability distributions the data was sampled from (28).

Let $T_i = (T_{i,1}, \ldots, T_{i,k})$, $i \in \{1, \ldots, N\}$, $k \in \mathbb{N}$ be sufficient statistics with respect to the deep latent variable model associated with a standard normal family of parameters $\lambda_i(X) = (\lambda_{i,1}(X), \ldots, \lambda_{i,k}(X))$, given a conditioning variable X. Importantly, the conditioning variable X is a proxy of the hidden confounder in our causal setup and functions λ_i are parameterized by LSTMs.

To illustrate, as per (27), let $k = 1$, set $T_i := T_{i,1}$, and let A be an invertible matrix. We can then recover $\mathbf{Z}$ related to the original $\mathbf{Z}^*$ as follows:

$$(T_1^*(Z_1^*), \ldots, T_n^*(Z_n^*)) = A(T_1(Z_1), \ldots, T_1(Z_1)). \tag{6}$$

This means we can estimate the original latent variables up to point-wise transformations T_i^*, T_i. In certain cases of non-sequential data, having A as a permutation matrix reduces the problem of indeterminacy of $\mathbf{Z}^*$ to finding the point-wise transformations of Z. This is due to Eq. (6) then becoming $T_i^*(Z_i^*) = T_{i'}(Z_{i'})$ for a permuted index i'.

Our deep latent variable model represented through parameters $\theta = (f, T_i, \lambda_i)$, as per (27), is

$$p_\theta(Y, W, Z|X) = p_f(Y, W|Z)p_{T_i,\lambda_i}(Z|X),$$

for a function f parameterized by an LSTM and sufficient statistic T_i with parameters λ_i, $i \in \{1, \ldots, N\}$. Some of the identifiability assumptions for a VAE as described in (27) are that function f is injective, and that sufficient statistic T_i is differentiable almost everywhere. Since our model is in principle a VAE, we rely on these assumptions and generate the synthetic data accordingly. We note, however, that in a dynamical-system setting this might not always suffice for the causal model to be identifiable and further theoretical work in this direction is required.

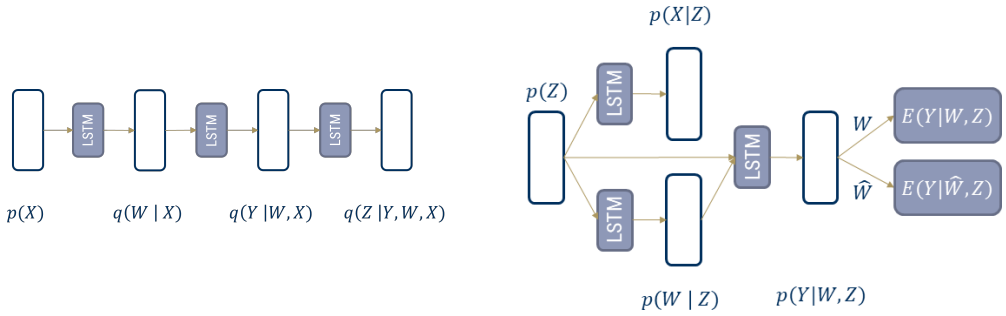

Figure 2: SCEVAE architecture. The encoder is shown on the left, and the decoder on the right.

3.3 Sequential CEVAE

Now we will introduce the Sequential Causal Effect Autoencoder (SCEVAE) (2). It is based on CEVAE (3) in which all distributions involved in the model and inference networks are modeled by LSTMs (29). Furthermore, it models $p(Y|W, X)$ sequentially, and introduces branching in the decoder as depicted on the right of Fig. 2. This branching allows it to compare factual and counterfactual causal effects, in order to estimate the cause-effect intensity after intervention on the non-binary sequential cause variable. The encoder of SCEVAE is depicted in Fig. 2 on the left.

We will now introduce the notation for sequential data and explain SCEVAE in more detail.

Let $X = \{x_t\}_{t=1}^N$, $W = \{w_t\}_{t=1}^N$, and $Y = \{y_t\}_{t=1}^N$ be the time series of proxy, cause variable, and effect , for $N \in \mathbb{N}$, respectively. The hidden confounder $Z = \{z_t\}_{t=1}^N$ is also time series for each $z_t \in \mathbb{R}^d$, where d is the dimension of Z in the latent space at each time step t. The conditional distributions of these variables used by SCEVAE framework are depicted in Fig. 2. The LSTMs are denoted by f_i, and each is parameterized by its own parameters ϕ_i, for $i \in \{1, 2, 3, 4\}$. For each time step t, we model x_t, w_t, and the prior of z_t as follows:

$$p(z_t) = \mathcal{N}(0, 1) \tag{7}$$

$$p(w_t|z_t) = \mathcal{N}(\mu_{w_t}, \sigma^2_{w_t}), \quad \mu_{w_t}, \sigma^2_{w_t} = f_1(z_t) \tag{8}$$

$$p(x_t|z_t) = \mathcal{N}(\mu_{x_t}, \sigma^2_{x_t}), \quad \mu_{x_t}, \sigma^2_{x_t} = f_2(z_t). \tag{9}$$

This means that each variable is full time series, since each individual time step is modeled by a Gaussian distribution. The LSTMs are not bidirectional, hence they do not have access to future values of the input variable. Using bidirectional LSTMs will be tackled in the future work.

When the proxy X contains binary components, SCEVAE models them according to $p(x_t|z_t) = \text{Bern}(\pi = \sigma(v(z_t)))$, for σ denoting a sigmoid function, and $v : \mathbb{R} \to \mathbb{R}$, a real-valued function parameterized by an LSTM.

The mean and variance of the effect y_t at each time step t are also parameterized by an LSTM:

$$p(y_t|w_t, z_t) = \mathcal{N}(\mu_{y_t}, \sigma^2_{y_t}), \quad \mu_{y_t}, \sigma^2_{y_t} = f_3(w_t, z_t) \tag{10}$$

We indicate that in contrast to CEVAE, SCEVAE does not fix variance in Eq. (10) but rather learn it from observational data. Since W is continuous and a time series, SCEVAE models this distribution at each time step, thus avoiding branching our method's architecture for different values of W, in contrast to CEVAE framework. SCEVAE approach is less restrictive, and can therefore be applied to a wider variety of real-world problems. The intervention it applies to W is setting it to Gaussian noise. In contrast, we use knockoffs (5) for this purpose, since they preserve the original distribution of the data as described in the following Subsection.

According to the DAG in Fig. 1, the posterior distribution of Z depends on X, Y, and W. SCEVAE thus approximates it by:

$$q(z_t|x_t, y_t, w_t) = \mathcal{N}(\mu_{z_t}, \sigma^2_{z_t}),\ \mu_{z_t}, \sigma_{z_t} = f_4(x_t, y_t, w_t). \tag{11}$$

For estimating the parameters of the auxiliary distribution of W and Y in Eq. (15), SCEVAE uses the following:

$$q(w_t|x_t) = \mathcal{N}(\mu^*_{w_t}, {\sigma^*}^2_{w_t}),\ \mu^*_{w_t}, {\sigma^*}^2_{w_t} = f_5(x_t) \tag{12}$$

$$q(y_t|x_t, w_t) = \mathcal{N}(\mu^*_{y_t}, {\sigma^*}^2_{y_t}), \mu^*_{y_t}, {\sigma^*}_{y_t} = f_6(x_t, w_t) \tag{13}$$

To make sure that variances are positive, SCEVAE uses a softplus activation function *softplus*$(u) = \ln(1 + e^u)$, for $u \in \mathbb{R}$ to the output of a given LSTM cell used to parameterize the variance. The regularization loss is weighted by $\lambda \in \mathbb{R}$ in order to obtain more stable effect predictions, so the variational lower bound becomes:

$$\hat{\mathcal{L}} = \sum_{i=1}^{N} \mathbb{E}_{q(z_t|x_t,w_t,y_t)}(\lambda \cdot (\log p(z_t) - \log q(z_t|x_t, w_t, y_t)) + \log p(x_t, w_t|z_t) + \log p(y_t|w_t, z_t)). \tag{14}$$

The variational lower bound used by SCEVAE is then:

$$\mathcal{F}_{\text{SCEVAE}} = \hat{\mathcal{L}} + \sum_{i=1}^{N} (\log q(w_t = w_t^*|x_t^*) + \log q(y_t = y_t^*|x_t^*, w_t^*)), \tag{15}$$

for x_t^*, w_t^*, y_t^* being the observed values for the input, intervention, and outcome variables in the training set.

3.4 Knockoffs

The idea of knockoffs was introduced by Barber and Candès (5) and it originates from the field of false discovery rate control in the setup of finding potential explanatory variables to an observed response variable. They are meant as a tool for estimating feature importance using conditional independence tests.

Let $u = \gamma \cdot f(\mathbf{Q}) + \eta$, be a predictive model for a vector of responses $\mathbf{u} \in \mathbb{R}^n$, an arbitrarily complex function f of a known design matrix $\mathbf{Q} \in \mathbb{R}^{n \times p}$ of potentially explanatory variables $(Q_1, \ldots, Q_p)$, an unknown vector of coefficients γ, and a Gaussian noise term $\eta \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. One can then generate a knockoff of any feature $Q_j \in \mathbf{Q}$ by first constructing a Gram matrix $\mathbf{\Sigma} = \mathbf{Q}^T\mathbf{Q}$, once all the features Q_j are normalized, so that $\Sigma_{j,j} = ||Q_j||_2^2 = 1$ for all $j = 1, \ldots, p$, and enforcing the following two conditions. First, the knockoffs of $\mathbf{Q}$, denoted as $\tilde{\mathbf{Q}} = (\tilde{Q}_1, \ldots, \tilde{Q}_p)$, are constructed to have the same covariance structure as $\mathbf{Q}$ by enforcing $\tilde{\mathbf{Q}}^T\tilde{\mathbf{Q}} = \mathbf{\Sigma}$ and $\mathbf{Q}^T\tilde{\mathbf{Q}} = \mathbf{\Sigma} - \text{diag}(\mathbf{s})$, for a p-dimensional non-negative vector $\mathbf{s}$. Moreover, the inter-variable correlations between different original and knockoff variables are enforced to be the same as those between the originals, that is, $Q_j^T\tilde{Q}_k = Q_j^T Q_k$ for all $j \neq k$.

We rely on the work by Romano et al. (6) to construct knockoffs for arbitrary and unspecified distributions of the observational data. We chose the semidefinite programming (SDP) knockoffs which are selected so that the original data and its knockoff are as decorrelated as possible.

4 Data

Causal link estimation is impossible to validate unless we know the underlying system dynamics. For this reason we use the following synthetic data to validate our approach with knockoff interventions.

4.1 Synthetic data

We create synthetic data with a nonlinear causal link according to the following equations. By z_t, x_t, w_t, $\hat{w}_t$, and y_t we denote the hidden confounder, its proxy, cause variable, knockoff $\tilde{w}_t$ intervention of the cause variable, and the effect variable at time step t, respectively.

$$z_t = a \cdot z_{t-1} + \epsilon_1 \tag{16}$$

$$x_t = \beta \cdot b_t \cdot \arctan(z_{t-\tau}) + s \cdot \epsilon_2 \tag{17}$$

$$w_t = c_1 \cdot w_{t-1} + c_2 \cdot z_t + \epsilon_3 \tag{18}$$

$$\hat{w}_t = do(w_t = \tilde{w}_t) \tag{19}$$

$$y_t = d_1 \cdot y_{t-1} + d_2 \cdot z_t + g \cdot w_t^3 + \epsilon_4 \tag{20}$$

To inspect how our approach behaves when causal effects are linear, we modify Eq. (20) as follows:

$$y_t = d_1 \cdot y_{t-1} + d_2 \cdot z_t + g \cdot w_t + \epsilon_4. \tag{21}$$

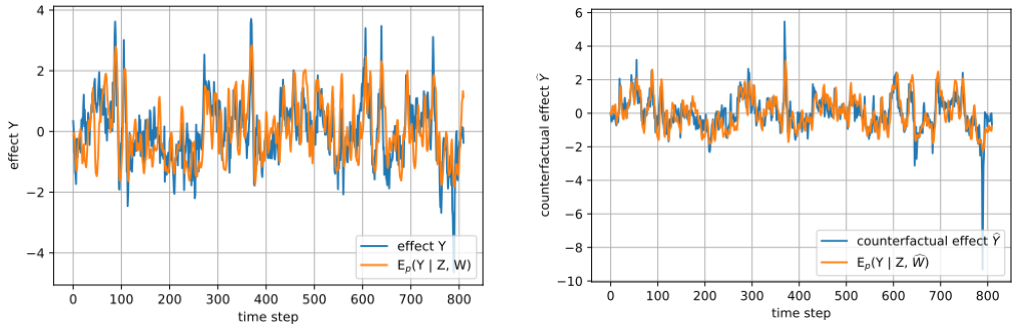

Figure 3: Left: Observational effect variable Y is shown in blue and the conditional expectation of Y given W and Z estimated by SCEVAE with knockoff intervention is shown in orange. Right: Counterfactual effect variable $\hat{Y}$ from Eq. (20) is shown in blue and the conditional expectation of Y given $\widehat{W}$ and Z estimated by SCEVAE with knockoff intervention is shown in orange.

In Eqs. (16)-(21), $a, \beta, c_i, d_i, g, h, s, \sigma^2, \sigma_j^2 \in \mathbb{R}$, $i = 1, 2$, $\tau \in \mathbb{N}$, and noise terms $\epsilon_j = \mathcal{N}(0, \sigma_j^2)$, for $j = 1, 2, 3, 4$, are mutually independent. The time-varying proxy parameter b_t is depicted in Fig. 4 (in Appendix) and generated by setting $b_t = \frac{2}{N-2}t$ for $t = 1, \ldots, \frac{N}{2} - 1$ and symmetrically setting $b_{N-t-1} := b_{t-1}$ for $t = \frac{N}{2}, \ldots, N$. The counterfactual outcome $\hat{y}_t$ for data with nonlinear and linear causal link is obtained by substituting w_t by $\hat{w}_t$ in Eq. (20) and Eq. (21), respectively.

The parameters used in our experiments are, as per (2), $a = 0.88$, $\beta = 0.5$, $s = 0.5$, $\tau = 100$, $c_1 = 0.6$, $c_2 = 0.2$, $d_1 = 0.4$, $d_2 = 0.8$, $g = 0.5$, $h = 0.5$, $\sigma^2 = 0.6$, $\sigma_2^2 = 1$, and $\sigma_1^2 = 0.5$, $\sigma_2^2 = 0.8$, $\sigma_3^2 = 0.7$, and $\sigma_4^2 = 0.5$. Furthermore, the hidden confounder and the cause variable are initialized with $z_0 = 0.1$ and $w_0 = 0.1$, whereas $y_0 = 0$.

4.2 Cloud and aerosol observation data

We demonstrate our method's applicability in real-world scenarios by applying it to cloud and aerosol observations dataset provided by Jesson et al. (30). It consists of the $1^\circ \times 1^\circ$ gridded version of the Moderate Resolution Imaging Spectroradiometer's measurements obtained twice per day at approximately 1 km $\times$ 1 km resolution from 2004 until 2019. Data stems from the Pacific Basin region off the coast of South America. As per (31), we use the daily average of all variables to downsample. As the outcome variable Y, we utilize cloud optical depth (COD) and as cause variable W the aerosol optical depth (AOD). In line with (31), we use the following meteorological variables as proxies. Namely, sea surface temperature (SST), estimated inversion strength (EIS), vertical motion at 500mb (w500), relative humidity at 700mb, 850mb, and 900mb (RH700, RH850, RH900). As intervention, we use the knockoff of the cause variable AOD. All variables are normalized before we use them as input for training SCEVAE.

5 Experiments

Here we expound our experimental setup. As per Trifunov et al. (2), we generate 1000-time-steps long multivariate time series and use the last 10% of each variable as test data without shuffling in order to preserve inter-temporal dependencies. The last 10% of training data is used for validation. We normalize the data by subtracting the mean of each variable and then dividing it by its standard deviation. We use batch size of 100 and randomly select that many consecutive time steps as batches of the training data. The optimizer used is Adam (32) with learning rate of 10^{-5}, and a weight decay of 10^{-3}. We model the hidden confounder as five-dimensional in the latent space in all but one experiment where we explicitly investigate the influence of its dimensionality to causal link estimation. The LSTMs of SCEVAE consist of two LSTM cells i.e. layers. Since the LSTMs that we use are not bidirectional, they can only look at the past time steps of a given time series. The hidden states and cell states of each LSTM have 32 dimensions. We set the regularization parameter λ from Eq. (14) to 0.1. For all experiments, we used GPUs of type GeForce RTX 2080 Ti.

5.1 Results

5.1.1 Synthetic data experiments

To demonstrate the efficacy of our approach by comparing its results to the ground truth causal link intensity, we apply it to synthetic data with either linear or nonlinear causal link between W and Y. Our method's reconstruction of the effect variable Y is shown in Fig. 3 on the left, whereas the reconstruction of the counterfactual effect variable $\hat{Y}$ is depicted in Fig. 3 on the right. We qualitatively observe that the reconstructions (orange) are quite close to the corresponding observed values (blue), accept in the case of outliers. These variables are generated according to Eq. (20) using W and $\widehat{W}$, respectively.

The quantitative results of SCEVAE's causal analysis with and without knockoff interventions are shown in Table 1. The causality scores used are ITE with a slight abuse of notation, denoting the difference between the ground truth and the predicted ITE as per Eq. (1), factual Root Mean Squared Error (RMSE), and counterfactual RMSE. The latter two scores measure the discrepancy between the observed Y and the estimated $\mathbb{E}(Y|W, Z)$ at each time step t, and the discrepancy between $\hat{Y}$ and $\mathbb{E}(Y|\widehat{W}, Z)$ at each time step t, respectively. In the linear case, as per Table 1, the estimated causal link using SCEVAE with knockoffs compared to the ground truth is ITE$= 0.44$ when proxy parameter b is time-varying and ITE$= 0.39$ for a fixed value of parameter $b = 0.95$. In the nonlinear case, for a time-varying parameter b ITE$= 0.5$, which is quite lower than in the case of SCEVAE without knockoff where ITE$= 0.66$. For a fixed $b = 0.95$ we obtain ITE$= 0.82$. We compare our results on synthetic data to those of Time Series Deconfounder (TSdeconf) (18) either without taking hidden confounding into account or when the substitutes for the hidden confounder are generated and used as proxies. We note that TSdeconf only outputs RMSE between the ground truth and the predicted outcome, but we generate the counterfactual RMSE by using $\hat{u}_t$ and $\hat{y}_t$ for training. Furthermore, since TSdeconf is not suitable for long time series, we generate 100 100-time-steps-long training samples for fairness. This is due to the fact that SCEVAE is trained on 100 100-long epochs randomly chosen from our 1000-long sequential variables.

We note that in the case of the nonlinear causal link, TSdeconf performs slightly better than our SCEVAE knockoff approach but with drastically higher variance, making our method much more stable. To obtain more stable results using TSdeconf, one would need to use much more training data. In the case of the linear causal link, our methods superior performance becomes clearer for both time-varying and fixed proxy parameter $b = 0.95$.

Table 1: Causal effect estimation metrics for linear and nonlinear causal link of synthetic data for $g = 0.2$. By "with" and "without" next to TSdeconf, we denote when the substitutes for the hidden confounder were used or if no confounders were taken into account, respectively. The results are averaged over five replications and shown with standard error.

Causal link	parameter b	Method	ITE	RMSE	RMSE cf
linear	time-varying	SCEVAE-knockoff	0.44 ± 0.06	0.79 ± 0.04	0.72 ± 0.04
		SCEVAE	0.58 ± 0.05	0.79 ± 0.04	1.04 ± 0.05
		TSdeconf (without)	/	0.87 ± 0.51	0.77 ± 0.31
		TSdeconf (with)	/	0.87 ± 0.52	0.77 ± 0.32
	0.95	SCEVAE-knockoff	0.39 ± 0.04	0.82 ± 0.04	0.76 ± 0.06
		TSdeconf (without)	/	0.88 ± 0.52	0.76 ± 0.3
		TSdeconf (with)	/	0.87 ± 0.5	0.76 ± 0.3
nonlinear	time-varying	SCEVAE-knockoff	0.5 ± 0.03	0.77 ± 0.02	0.72 ± 0.03
		SCEVAE	0.66 ± 0.02	0.73 ± 0.02	0.96 ± 0.02
		TSdeconf (without)	/	0.74 ± 0.32	0.57 ± 0.22
		TSdeconf (with)	/	0.74 ± 0.32	0.54 ± 0.2
	0.95	SCEVAE-knockoff	0.82 ± 0.3	0.78 ± 0.04	1 ± 0.42
		TSdeconf (without)	/	0.74 ± 0.32	0.53 ± 0.24
		TSdeconf (with)	/	0.74 ± 0.31	0.58 ± 0.25

The counterfactual RMSE metric shows that SCEVAE with knockoff intervention generates better counterfactual forecast than SCEVAE with standard normal intervention. This is expected since $P(W|Z)$ is more similar to $P(\widehat{W}|Z)$ in the case when $\widehat{W}$ is a knockoff of W, sharing its distribution. One can also note that SCEVAE with knockoff intervention performs better when parameter b is not time-varying in the nonlinear case, but that the opposite is true in the linear case.

In Fig. 6 in Appendix we show how dimensionality of Z in the latent space and the confounding coefficient d_2 influence the factual and counterfactual outcome's reconstruction. We tested latent dimensions $D_Z \in \{1, 5, 10, 20\}$, and coefficients $d_2 \in \{0.8, 1, 1.2\}$. We observe that RMSE values in both factual and counterfactual cases are lowest when $d_2 = 0.8$ and tend to increase as the rate of hidden confounding d_2 increases regardless of D_Z. As the dimensionality of Z increases, the uncertainty of the prediction becomes higher.

The factual reconstruction RMSE during training and test can be found in Appendix Fig. 5. We note that by increasing the number of latent dimensions the standard deviation of the prediction increases.

5.1.2 Cloud and aerosol data experiments

In the experiments on real aerosol-cloud-climate observations, we demonstrate the importance of choosing good proxy variables and how SCEVAE with knockoff intervention improves the causal link estimation results. Since now we do not have the ground truth causal link intensity values, we cannot use the ITE error. Instead, we use ATE of the predicted factual and counterfactual outcome variables Y and $\hat{Y}$. In all experiments from Table 2 we use COD as outcome and AOD as intervention variables, but we vary the choice of proxy variable X. In case of meteorological proxy, we set it to be multivariate including SST, EIS, w500, RH700, RH850, and RH900 as per (31). In the experiments where proxy is indicated as *uniform*, we set it univariately to uniform $\mathcal{U}(0, 1)$ noise.

Table 2: Causal effect estimation metrics for cloud-aerosol dataset when using COD as the outcome variable and AOD as intervention variable. In the unmarked row we used meteorological proxies EIS, SST, w500, RH700, RH850, and RH900. In the row marked by * we set proxy X to uniform $\mathcal{U}(0, 1)$ noise. The values are averaged over five replications and shown with standard error.

Method	Proxy	ATE train	ATE test	RMSE
SCEVAE-knockoff	meteorological	0.21 ± 0.13	0.14 ± 0.09	0.95 ± 0.13
	uniform	0.13 ± 0.16	0.38 ± 0.33	0.74 ± 0.11
SCEVAE	meteorological	0.06 ± 0.05	0.09 ± 0.08	0.72 ± 0.001
	uniform	0.21 ± 0.18	0.24 ± 0.14	0.79 ± 0.14

We note that using meteorological proxies yields lower variance of both training and test ATE in contrast to the case where we used uniform noise as the proxy. Moreover, we note that the true causal link between COD and AOD measured by ATE is stronger when we use SCEVAE with knockoff intervention than without it. Furthermore, when using knockoff intervention the causal link intensity between COD and AOD is stronger during training when meteorological proxies are employed than in the case of setting X to uniform noise. This is opposite during test.

6 Conclusion

Causal link intensity estimation is a challenging task, especially in the presence of latent confounders. In this paper we proposed to use SCEVAE with knockoff intervention for this purpose. We achieved more stable results than the time series deconfounding benchmark TSdeconf and were able to estimate the confounded causal link intensity more accurately than SCEVAE with standard normal interventions. This was attained on synthetic data with both linear and nonlinear causal links. In addition, we observed that using knockoff interventions reduces the counterfactual RMSE in comparison to SCEVAE without knockoff intervention since the distribution of the knockoff agrees with the learned distribution of the cause variable. Moreover, through the experiments on real cloud-aerosol observational data, we indicated our method's applicability to real-world problems and illustrated how the use of proxy contributes to better causal link intensity estimation.

References

[1] J. Runge, S. Bathiany, E. Bollt, *et al.*, "Inferring causation from time series of earth system sciences," *Nature Communications*, vol. 10, p. 2553, 2019.

[2] V. T. Trifunov, M. Shadaydeh, and J. Denzler, "Sequential causal effect variational autoencoder: Time series causal link estimation under hidden confounding," *arXiv:2209.11497*, 2022.

[3] C. Louizos, U. Shalit, J. Mooij, D. Sontag, R. Zemel, and M. Welling, "Causal effect inference with deep latent-variable models," in *Advances in Neural Information Processing Systems 30*, pp. 6446–6456, Curran Associates Inc., 2017.

[4] S. Samarasinghe, E. A. Barnes, and I. Ebert-Uphoff, "Causal discovery in the presence of confounding latent variables for climate science," 2018.

[5] R. F. Barber and E. J. Candès, "Controlling the false discovery rate via knockoffs," *The Annals of Statistics*, vol. 43, no. 5, 2015.

[6] Y. Romano, M. Sesia, and E. Candès, "Deep knockoffs," *Journal of the American Statistical Association*, vol. 115, no. 532, pp. 1861–1872, 2020.

[7] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd ed., 2009.

[8] A. M. Alaa and M. van der Schaar, "Bayesian inference of individualized treatment effects using multi-task gaussian processes," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[9] K. Jiang and Y. Ning, "Treatment effect estimation with unobserved and heterogeneous confounding variables," 2022.

[10] Z. Qian, Y. Zhang, I. Bica, A. Wood, and M. van der Schaar, "Synctwin: Treatment effect estimation with longitudinal outcomes," in *Advances in Neural Information Processing Systems*, vol. 34, pp. 3178–3190, Curran Associates, Inc., 2021.

[11] L. Nie, M. Ye, Q. Liu, and D. Nicolae, "Vcnet and functional targeted regularization for learning causal effects of continuous treatments," 2021.

[12] T. Wolf, L. Debut, V. Sanh, *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Association for Computational Linguistics, 2020.

[13] Y. Zhang, H. Zhang, Z. C. Lipton, L. E. Li, and E. P. Xing, "Exploring transformer backbones for heterogeneous treatment effect estimation," *CoRR*, vol. abs/2202.01336, 2022.

[14] V. Melnychuk, D. Frauen, and S. Feuerriegel, "Causal transformer for estimating counterfactual outcomes," 2022.

[15] S. Rissanen and P. Marttinen, "A critical look at the consistency of causal estimation with deep latent variable models," in *Advances in Neural Information Processing Systems*, vol. 34, pp. 4207–4217, Curran Associates, Inc., 2021.

[16] D. J. Im, K. Cho, and N. Razavian, "Causal effect variational autoencoder with uniform treatment," *CoRR*, vol. abs/2111.08656, 2021.

[17] V. T. Trifunov, M. Shadaydeh, J. Runge, V. Eyring, M. Reichstein, and J. Denzler, "Nonlinear causal link estimation under hidden confounding with an application to time-series anomaly detection," in *German Conference on Pattern Recognition (GCPR)*, pp. 261–273, Springer-Verlag, 2019.

[18] I. Bica, A. M. Alaa, and M. van der Schaar, "Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders," 2019.

[19] T. Hatt and S. Feuerriegel, "Sequential deconfounding for causal inference with unobserved confounders," vol. abs/2104.09323, 2021.

[20] O.-I. Popescu, M. Shadaydeh, and J. Denzler, "Counterfactual generation with knockoffs," *arXiv:2102.00951*, 2021.

[21] W. Ahmad, M. Shadaydeh, and J. Denzler, "Causal inference in non-linear time-series using deep networks and knockoff counterfactuals," in *20th IEEE International Conference on Machine Learning and Applications*, 09 2021.

[22] W. Ahmad, M. Shadaydeh, and J. Denzler, "Causal discovery using model invariance through knockoff interventions," in *ICML Workshop on Spurious Correlations, Invariance and Stability (ICML-WS)*, 2022.

[23] Z. Yin and P. Barucca, "Deep recurrent modelling of granger causality with latent confounding," *Expert Systems with Applications*, vol. 207, p. 118036, 2022.

[24] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.

[25] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014.

[26] U. Shalit, F. Johansson, and D. Sontag, "Estimating individual treatment effect: Generalization bounds and algorithms," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, p. 3076–3085, JMLR.org, 2017.

[27] I. Khemakhem, D. P. Kingma, R. P. Monti, and A. Hyvärinen, "Variational autoencoders and nonlinear ica: A unifying framework," in *AISTATS 2020 - 23rd International Conference on Artificial Intelligence and Statistics*, 2020.

[28] R. A. Fisher, "A mathematical Examination of the Methods of determining the Accuracy of Observation by the Mean Error, and by the Mean Square Error," *Monthly Notices of the Royal Astronomical Society*, vol. 80, no. 8, pp. 758–770, 1920.

[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[30] A. Jesson, P. Manshausen, A. Douglas, *et al.*, "Using non-linear causal models to study aerosol-cloud interactions in the southeast pacific," 2021.

[31] A. Jesson, A. Douglas, P. Manshausen, *et al.*, "Scalable sensitivity and uncertainty analyses for causal-effect estimates of continuous-valued interventions," *arXiv preprint arXiv:2204.10022v3*, 2022.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2015.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] We have not included the code, but we provided the equations through which one can recreate the synthetic data, and the real data we used is publicly available. The instructions to reproduce the main experimental results are provided in the text in Section 3.3 and Section 5.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We included a new synthetic dataset model.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

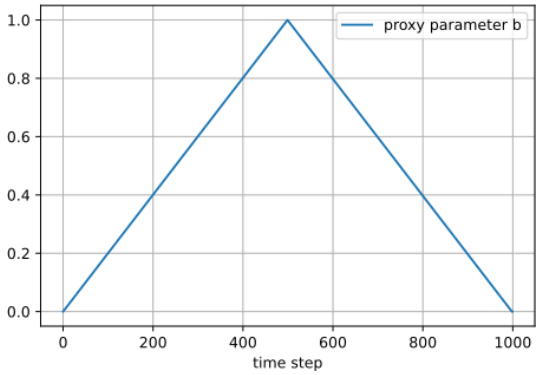

Figure 4: Time-varying proxy parameter b_t over 1000 time steps.

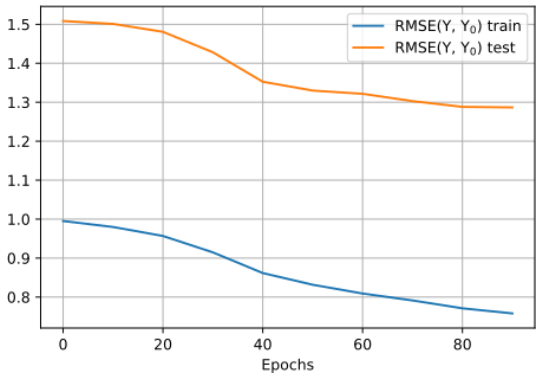

Figure 5: RMSE during training (blue) and test (orange) for the synthetic data with nonlinear causal link and time-varying proxy parameter b.

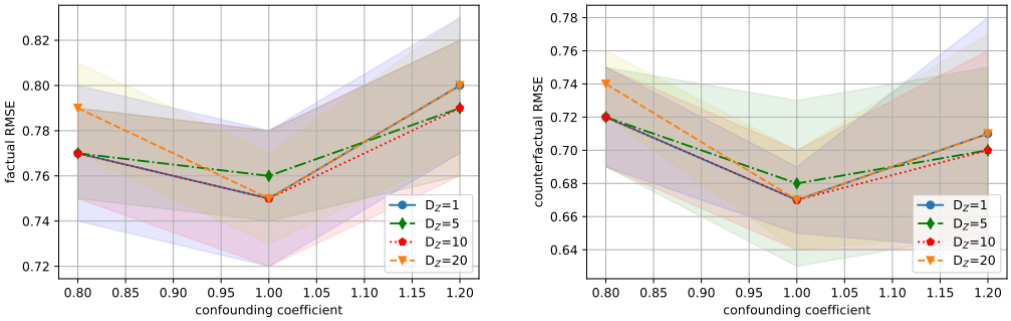

Figure 6: Influence of latent dimension and coefficient of the latent confounder to the outcome forecast. Factual RMSE is shown on the left and counterfactual RMSE on the right with standard deviation after five replications.