Time Series Causal Link Estimation under Hidden Confounding using Knockoff Interventions

Violeta Teodora Trifunov
Department for Mathematics and Computer Science, Computer Vision Group
Friedrich Schiller University Jena
Jena, Germany
violetateodora.trifunov@uni-jena.de

Maha Shadaydeh
Department for Mathematics and Computer Science, Computer Vision Group
Friedrich Schiller University Jena
Jena, Germany

Joachim Denzler
Department for Mathematics and Computer Science, Computer Vision Group
Friedrich Schiller University Jena
Jena, Germany

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. Knockoff variables have the same distribution as the originals and preserve the correlation to other variables. This allows for counterfactuals that are more faithful to the observational distribution. 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), (2). We propose a novel method for estimating both linear and nonlinear causal effects in time series under hidden confounding while performing Knockoff (3) interventions, inspired by the Causal Effect Variational Autoencoder (CEVAE) (4). The proposed approach, as will be explained further on, allows for less biased causality analysis of non-stationary sequential data and counterfactuals that are more faithful to the distribution of observational data. A

A causal view on dynamical systems workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

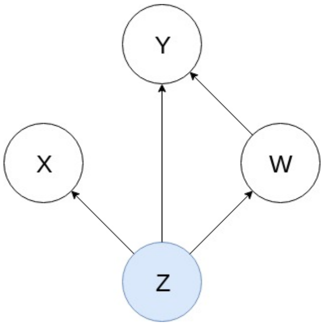

Figure 1: Causal DAG with observed variables $\{W, X, Y\}$ and latent variable $\{Z\}$.

hidden confounder is a variable that influences both the cause and the effect, and its presence can lead to false relationships which may be misconceived as causal.

CEVAE is 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 (5). 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 the identifiability theory in Section 3.2.

Knockoff filters were first introduced by Barber and Candès in 2015 (3) 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.

Due to its highly practicable properties for counterfactual analysis (7), as proposed by (8), we generate a Knockoff (3) of the cause variable W as a way of intervention to estimate the causal link intensity using our approach. 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. To highlight the advantage of Knockoff intervention, we compare it to using Gaussian noise as intervention variable $\widehat{W}$. It is independent of the hidden confounder Z and therefore removes the link between W and Z as required by do-calculus (7). However, Gaussian noise intervention does not preserve the distribution of the original cause variable W, which is learned during the training of the neural networks within our method and expected for inference. This leads to suboptimal performance in contrast to Knockoff interventions.

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.

To address these issues, we engineered our approach to transcend the CEVAE methodology and become applicable to complicated non-stationary sequential data with the help of Long Short-Term Memory (LSTM) (9) recurrent neural networks. Moreover, we allow for the causal effect estimation to be done on a single time series example of each observed variable. This means that one does not need to have multiple realizations of the variables on their disposal which is an advantage in many fields such as finance and climate science where each variable is observed only once per time step.

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 associated work that utilizes Knockoffs. In Section 3, we introduce the CEVAE framework, define and discuss causal identifiability, the assumptions under which using SCEVAE is justified, as well as a detailed description of our method, and introduce Knockoffs themselves in more detail. Section 4 describes the data we use, and Section 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 a causal effect estimation method under hidden confounding for sequential data based on CEVAE and utilize it with Knockoff interventions.

2 Related Work

Treatment effect estimation in the static setting is a very active field of research (10), (11), (12). Under the assumption of no hidden confounding, methods such as VCNet (13) and its extensions by state-of-the-art transformer networks (14) were recently applied to the task of continuous treatment estimation by Zhang et al. (15) and Melnychuk et al. (16).

In a more realistic scenario when there are unobserved confounders present, Causal Effect Variational Autoencoder (CEVAE) (4) 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 abovementioned transformer-based treatment estimation methods, CEVAE focuses on binary treatment. Research endeavors such as those by Rissanen and Marttinen (17), Im et al. (18), and Trifunov et al. (19) either extend CEVAE to a uniform or continuous treatment, or analyse this deep latent variable model's capabilities. However, until now, to the best of our knowledge, CEVAE has never been applied to time series.

Taking a different approach towards treatment effect estimation, Bica et al. (20) and Hatt and Feuerriegel (21) propose deconfounding methods for sequential data using neural networks. Specifically, in the work by Bica et al. (20), 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 ours, 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. (21), 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, 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 (5).

In the recent years, due to their distribution- and correlation-preserving properties, Knockoffs (3), (6) were innovatively applied for counterfactual analysis of images (22), and time series (23), (8). Ahmad et al. (23) 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 (24) 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 (25) 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, as well as lower counterfactual prediction error as will be discussed 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) (4). It is a deep learning method based on a VAE (26) and a TARnet (27). 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:

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

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

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. (4).

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.

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. (28) 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

$$p_\theta(Y) = p_{\hat{\theta}}(Y) \Rightarrow \theta \sim \hat{\theta} \tag{3}$$

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 $i \in \{1, \dots, N\}$ in the form of sufficient statistics, and invertible matrix A (28).

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 (29).

Let $T_i = (T_{i,1}, \dots, T_{i,k})$, $i \in \{1, \dots, N\}$, $k \in \mathbb{N}$ be sufficient statistics with respect to the deep latent variable model associated with a standard normal family of parameters $\lambda_i(X) = (\lambda_{i,1}(X), \dots, \lambda_{i,k}(X))$, 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 (28), 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:

$$(T_1^*(Z_1^*), \dots, T_n^*(Z_n^*)) = A(T_1(Z_1), \dots, T_1(Z_1)). \tag{4}$$

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. (4) 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 (28), is

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

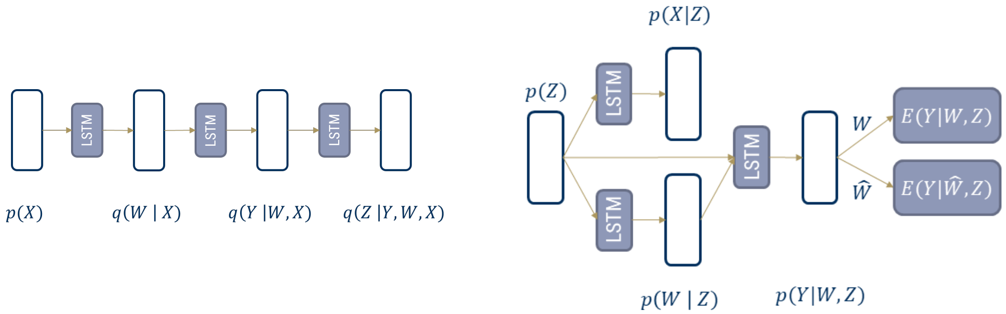

Figure 2: SCEVAE architecture. The encoder is shown on the left, and the decoder with branching to estimate the factual and counterfactual outcomes on the right.

for a function f parameterized by an LSTM and sufficient statistic T_i with parameters λ_i, $i \in \{1, \dots, N\}$. Some of the identifiability assumptions for a VAE as described in (28) 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.

3.3 Sequential CEVAE

We propose a novel method for time series causal link estimation under hidden confounding. It is predicated on CEVAE, but differs significantly in the architecture and relies on LSTMs to model the involved probability distributions. Namely, we do not use a TARnet to estimate $p(Y|X, w^0)$ and $p(Y|X, w^1)$, specialized for binary treatment, but rather model $p(Y|X, W)$ sequentially, and introduce branching in the decoder as depicted on the right of Fig. 2. This branching allows us to compare factual and counterfactual causal effects, in order to estimate the cause-effect intensity after intervention on the non-binary sequential cause variable.

We will now introduce the notation for sequential data and explain the proposed method in full 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 our method's 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:

$$p(z_t) = \mathcal{N}(0, 1) \tag{5}$$

$$p(w_t|z_t) = \mathcal{N}(\mu_{w_t}, \sigma^2_{w_t}), \quad \mu_{w_t}, \sigma^2_{w_t} = f_1(z_t) \tag{6}$$

$$p(x_t|z_t) = \mathcal{N}(\mu_{x_t}, \sigma^2_{x_t}), \quad \mu_{x_t}, \sigma^2_{x_t} = f_2(z_t). \tag{7}$$

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

When proxy X contains binary components, we model them as $p(x_t|z_t) = \text{Bern}(\pi = \sigma(v(z_t)))$, for the sigmoid function σ, 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:

$$p(y_t|w_t, z_t) = \mathcal{N}(\mu_{y_t}, \sigma^2_{y_t}), \quad \mu_{y_t}, \sigma^2_{y_t} = f_3(w_t, z_t) \tag{8}$$

We indicate that in contrast to CEVAE, we do not fix variance in Eq. (8) but rather learn it from observational data. Our approach is less restrictive in the sense of its compatibility with continuous, sequential cause variable, and can therefore be applied to a wider variety of real-world problems. We use a Knockoff (3) of W as intervention since it preserves the original distribution of the data. This is of particular importance for creating counterfactuals because the trained neural network anticipates test and training data stemming from the same distribution. For comparison, we also apply Gaussian noise as intervention on W which removes the link to the hidden confounder but does not preserve the original distribution of W. One limitation of our approach is that we assume that, within the processed window, data is stationary and sufficient to produce good Knockoffs.

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

$$q(z_t|x_t, y_t, w_t) = \mathcal{N}(\mu_{z_t}, \sigma^2_{z_t}),\ \mu_{z_t}, \sigma^2_{z_t} = f_4(x_t, y_t, w_t). \tag{9}$$

To estimate the parameters of the auxiliary distribution of W and Y in Eq. (13), we use the following:

$$q(w_t|x_t) = \mathcal{N}(\mu^*_{w_t}, {\sigma^*_{w_t}}^2),\ \mu^*_{w_t}, {\sigma^*}^2_{w_t} = f_5(x_t) \tag{10}$$

$$q(y_t|x_t, w_t) = \mathcal{N}(\mu^*_{y_t}, {\sigma^*}^2_{y_t}), \mu^*_{y_t}, {\sigma^*}^2_{y_t} = f_6(x_t, w_t) \tag{11}$$

To insure that variances are positive, we apply 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.

We weigh the regularization loss by $\lambda \in \mathbb{R}$ in order to obtain more stable effect predictions, so the variational lower bound involving all modeled distributions is:

$$\hat{\mathcal{L}} = \sum_{i=1}^{N} \mathbb{E}_{q(z_t|x_t,w_t,y_t)}(\lambda(\log p(z_t) - \log q(z_t|x_t, w_t, y_t)) + \log p(x_t, w_t|z_t) + \log p(y_t|w_t, z_t)). \tag{12}$$

Similarly to CEVAE, 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_t and y_t for new samples. The variational lower bound used as an objective for both the inference and the model networks of our method is then:

$$\mathcal{F}_{\text{SCEVAE}} = \hat{\mathcal{L}} + \sum_{i=1}^{N} \big(\log q(w_t = w^*_t|x^*_t) + \log q(y_t = y^*_t|x^*_t, w^*_t)\big), \tag{13}$$

for x^*_t, w^*_t, y^*_t being the observed values for the input, intervention, and outcome variables in the training set.

Our method's encoder and decoder are depicted in Fig. 2. We call this novel approach Sequential Causal Effect Variational Autoencoder (SCEVAE, pronounced /see-VAE/).

3.4 Knockoffs

The idea of Knockoffs was introduced by Barber and Candès (3) 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 but have recently been applied as intervention variables for causal discovery (8).

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 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, $Q^T_j\tilde{Q_k} = Q^T_jQ_k$ for all $j \neq k$. This condition is known as *exchangeability*.

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

The Knockoffs are generated by using Gaussian mixture models as proposed by Gimenez et al. (30). First, the mixture assignment variable is sampled from the posterior distribution. The Knockoffs are then sampled from the conditional distribution given the original variables and the sampled mixture assignment such that the exchangeability condition is fulfilled (30).

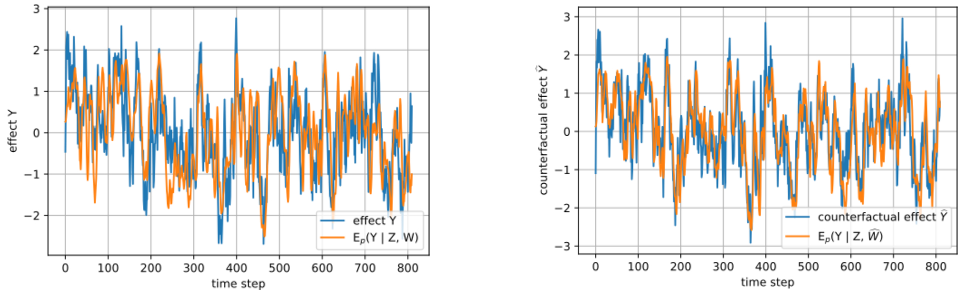

Figure 3: Left: Observational effect variable Y from Eq. (17) (blue) and the conditional expectation of Y given Z and W (orange). Right: Counterfactual $\hat{Y}$ (blue) of the effect variable Y, and the conditional expectation of Y given Z and $\widehat{W}$ (orange). Conditional expectations are estimated by SCEVAE with Knockoff intervention.

4 Data

4.1 Synthetic data

Causal link estimation is impossible to validate unless we know the underlying system dynamics. For this reason we generate synthetic data. It contains a nonlinear causal link, all variables abide by the relations from DAG in Fig. 1 and form the structural causal model 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, and the effect variable at time step t, respectively.

$$z_t = a \cdot z_{t-1} + \epsilon_{1t} \tag{14}$$

$$x_t = b_t \cdot \tanh(z_{t-\tau}) + s \cdot \epsilon_{2t} \tag{15}$$

$$w_t = c_1 \cdot w_{t-1} + c_2 \cdot z_t + \epsilon_{3t} \tag{16}$$

$$y_t = d_1 \cdot y_{t-1} + d_2 \cdot z_t + g \cdot e^{h \cdot w_t} + \epsilon_{4t} \tag{17}$$

In the case of Knockoff intervention $\tilde{w}_t$, the cause variable becomes $\hat{w}_t = do(w_t = \tilde{w}_t)$, and in the Gaussian noise intervention case we have $\hat{w}_t \sim \mathcal{N}(0, \sigma^2)$, for $\sigma^2 \in \mathbb{R}$. To asses how our method performs for estimating linear causal effects, we alter Eq. (17) as follows:

$$y_t = d_1 \cdot y_{t-1} + d_2 \cdot z_t + g \cdot w_t + \epsilon_{4t} \tag{18}$$

In Eqs. (14)-(18), $a, c_i, d_i, g, h, s, \sigma_j^2 \in \mathbb{R}$, $i = 1, 2$, $\tau \in \mathbb{N}$, and noise terms $\epsilon_j = \mathcal{N}(0, \sigma_j^2)$, for $j = 1, 2, 3, 4$, are mutually independent. By e we denote the Euler's number. The time-varying proxy parameter b_t (Fig. 4, Appendix A) is generated by setting $b_t = \frac{2}{N-2}t$ for $t = 1, \ldots, \frac{N}{2} - 1$ and symmetrically $b_{N-t-1} := b_{t-1}$ 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. (17) and Eq. (18), respectively. The parameters used in our experiments are $a = 0.88$, $s = 0.5$, $\tau = 100$, $c_1 = 0.6$, $c_2 = 0.2$, $d_1 = 0.4$, $d_2 = 0.8$, $g = 0.5$, $h = 0.5$, $\sigma^2 = 0.6$, $\sigma_2^2 = 1$, and $\sigma_1^2 = 0.5$, $\sigma_2^2 = 0.8$, $\sigma_3^2 = 0.7$, and $\sigma_4^2 = 0.5$. 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 performance in real-world scenarios by applying it to cloud and aerosol observations dataset provided by Jesson et al. (31). It consists of the 1° × 1° gridded version of the Moderate Resolution Imaging Spectroradiometer's measurements obtained twice per day at approximately 1 km × 1 km resolution from 2004 until 2019. Data stems from the Pacific Basin region off the coast of South America. As per (32), 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 (32), 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). All variables are normalized before we use them as input for training SCEVAE.

5 Experiments

Here we expound our experimental setup. We generate 1000-time-step-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 (33) 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. (12) 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 proposed method 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 and its counterfactual is $\hat{Y}$ is depicted in Fig. 3. We qualitatively observe that the reconstructions (orange) are closely comparable to the corresponding observed values (blue). These variables are generated according to Eq. (17) using W and $\widehat{W}$, respectively.

The quantitative results of SCEVAE's causal analysis are shown in Table 1. The causality scores used are RMSE_{ITE}, denoting the RMSE between the ground truth and the predicted ITE as per Eq. (1), factual RMSE_Y, and counterfactual $\text{RMSE}_{\hat{Y}}$. 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. By "SCEVAE" in Table 1, we denote our method with Gaussian noise intervention on W, whereas "SCEVAE-Knockoff" indicates that we used the Knockoff of W as intervention $\widehat{W}$.

Table 1: Causal Effect Estimation Error Metrics for Linear and Nonlinear Causal Link of Synthetic Data for $g = 0.5$. The results are averaged over five replications and shown with standard error. Lower is better.

Causal link	parameter b	Method	RMSE_{ITE}	RMSE_Y	$\text{RMSE}_{\hat{Y}}$
linear	time-varying	SCEVAE	0.52 ± 0.04	0.74 ± 0.06	0.99 ± 0.03
		SCEVAE-Knockoff	0.34 ± 0.03	0.74 ± 0.01	0.63 ± 0.01
		TSdeconf (without)	/	0.75 ± 0.28	0.93 ± 0.41
		TSdeconf (with)	/	0.75 ± 0.28	0.93 ± 0.41
	0.95	SCEVAE	0.57 ± 0.06	0.72 ± 0.02	0.97 ± 0.03
		SCEVAE-Knockoff	0.32 ± 0.01	0.76 ± 0.08	0.65 ± 0.07
		TSdeconf (without)	/	0.76 ± 0.29	0.93 ± 0.4
		TSdeconf (with)	/	0.75 ± 0.28	0.93 ± 0.41
nonlinear	time-varying	SCEVAE	0.56 ± 0.05	0.74 ± 0.03	0.99 ± 0.03
		SCEVAE-Knockoff	0.44 ± 0.03	0.81 ± 0.03	0.7 ± 0.03
		TSdeconf (without)	/	0.85 ± 0.34	0.99 ± 0.47
		TSdeconf (with)	/	0.84 ± 0.33	0.98 ± 0.46
	0.95	SCEVAE	0.59 ± 0.1	0.79 ± 0.03	1.01 ± 0.01
		SCEVAE-Knockoff	0.43 ± 0.03	0.8 ± 0.03	0.7 ± 0.03
		TSdeconf (without)	/	0.84 ± 0.33	0.98 ± 0.46
		TSdeconf (with)	/	0.84 ± 0.33	0.98 ± 0.45

According to the results from Table 1, SCEVAE performs better in the linear than in the nonlinear causal link case, which illustrates the higher complexity of estimating nonlinear causal links. We note that the methods works equally well for both fixed and time-varying values of the parameter b. Furthermore, in Table 1, we can clearly see how using Knockoffs improves the causal link intensity estimation by remarkably lower RMSE_{ITE} metric, as well as by inducing lower variance, making this a more reliable intervention choice. In addition, it is especially interesting to note that the counterfactual RMSE is much lower in comparison to other methods which do not use Knockoff interventions. This is due to the fact that Knockoff $\widehat{W}$ has the same distribution as W, and therefore produces a good reconstruction of the counterfactual outcome $\hat{Y}$ when we sample it from the learned estimate of $p(Y|W, Z)$.

We compare our results on synthetic data to those of Time Series Deconfounder (TSdeconf) (20) either without taking hidden confounding into account or when the substitutes for the hidden confounder are generated and used as proxies, in Table 1 denoted by "TSdeconf (without)" and "TSdeconf (with)", respectively. We set the confounding parameter of TSdeconf to $\gamma = 0.8$, and the number of substitute, as well as the simulated hidden confounders to one. We note that TSdeconf only outputs RMSE between the ground truth and the predicted factual or counterfactual outcome, so these are the main comparison metrics. For obtaining the factual RMSE_T via TSdeconf we input X as covariates, W as treatment, and Y as outcome. To obtain the counterfactual $\text{RMSE}_{\hat{Y}}$, we also use X as covariates, but $\widehat{W}$, and $\hat{Y}$ as treatment and outcome, respectively. Furthermore, since TSdeconf is not suitable for long time series, we generate 100 100-time-step-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-time-step-long sequential variables.

We note that in the case of the linear causal link, TSdeconf performs almost as well as our method but with much higher variance, making our method considerably more stable. To obtain stable results using TSdeconf, one would need much larger amount of training data. In the case of the nonlinear causal link, our method's superior performance becomes even clearer for both time-varying, and fixed proxy parameter $b = 0.95$.

The factual RMSE_T of the outcome reconstruction during training and test can be found in Fig. 5 (Appendix A). In Fig. 6 (Appendix A) 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 $d_2 \in \{0.8, 1, 1.2, 1.6\}$. We observe that RMSE values in both factual and counterfactual cases are lowest when $d_2 = 0.8$ and have an upward trend 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.

5.1.2 Cloud and aerosol data experiments

In the experiments on real aerosol-cloud-climate observations, where we use COD as the effect and AOD as the cause variable, we demonstrate the importance of choosing suitable proxies. Since now we do not have the ground truth causal link intensity values, we cannot use RMSE_{ITE}. Instead, we use ATE metric as per Eq. (2) of the predicted factual and counterfactual outcome variables Y and $\hat{Y}$, respectively. Moreover, we note that intervention on real data is often not feasible in practice, but that by our method's way of intervention we are able to counterfactually analyse the desired causal link.

Table 2: Causal Effect Estimation Error Metrics for Cloud-Aerosol Dataset when using COD as Outcome Variable and AOD as Intervention Variable. The results are averaged over five replications and shown with standard error.

Method	Proxy	ATE train	ATE test	RMSE_Y
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
SCEVAE-Knockoff	meteorological	0.03 ± 0.02	0.04 ± 0.02	0.78 ± 0.02
	uniform	0.29 ± 0.08	0.37 ± 0.06	0.83 ± 0.03

In case of meteorological proxy, we set it to be multivariate including SST, EIS, w500, RH700, RH850, and RH900 as per Jesson et al. (32). Whereas when the proxy is indicated as *uniform*, we set it univariately to $\mathcal{U}(0, 1)$. The results of varying the choice of proxy X are shown in Table 2 for SCEVAE with standard normal, as well as with Knockoff intervention.

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, SCEVAE with Knockoff intervention yields lower RMSE_Y with the use of meteorological proxies. This implies the importance of the appropriate proxy choice.

Furthermore, when using meteorological proxies the causal link intensity between COD and AOD is lower than in the case we set X to uniform noise during both training and test. Similar findings, i.e. that aerosol has limited impact on cloud depth, were reported by Jesson et al. (32) for a comparable rate of confounding. This further strengthens our claim that choosing suitable proxies is crucial for correctly estimating causal link intensity under hidden confounding as it may contribute to identifiability.

6 Conclusion

Causal link intensity estimation is a challenging task, especially in the presence of latent confounders. In this paper we introduced SCEVAE, a novel deep learning method for time series causality analysis under hidden confounding to tackle this problem in sequential data with Knockoff interventions. It is inspired by the CEVAE framework, but applicable to complex non-stationary time series through the use of LSTMs and fundamental architectural novelty. Our method allows for single-variable causality analysis instead of using many independent samples for training. We achieved better and more stable results than the time series deconfounding benchmark TSdeconf. Moreover, we showed that estimating the confounded causal link intensity can be done more accurately with Knockoff rather than using 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 $\text{RMSE}_{\hat{Y}}$ 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 potential applicability to real-world problems and illustrated how the use of meaningful proxies contributes to its identifiability.

Acknowledgements

Funding for this study was provided by the German Research Foundation (DFG) individual research grant SH 1682/1-1 and the European Research Council (ERC) Synergy Grant "Understanding and Modelling the Earth System with Machine Learning (USMILE)" under the Horizon 2020 research and innovation programme (Grant agreement No. 855187).

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] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning, MIT Press, 2017.

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

[4] 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.

[5] S. Samarasinghe, E. A. Barnes, and I. Ebert-Uphoff, "Causal discovery in the presence of confounding latent variables for climate science," in *8th Internvational Workshop on Climate Informatics*, 2018.

[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] 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.

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

[10] 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.

[11] K. Jiang and Y. Ning, "Treatment effect estimation with unobserved and heterogeneous confounding variables," in *10.48550/arXiv.2207.14439*, 2022.

[12] 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.

[13] L. Nie, M. Ye, Q. Liu, and D. Nicolae, "Vcnet and functional targeted regularization for learning causal effects of continuous treatments," in *9th International Conference on Learning Representations, (ICLR)*, 2021.

[14] 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.

[15] 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.

[16] V. Melnychuk, D. Frauen, and S. Feuerriegel, "Causal transformer for estimating counterfactual outcomes," in *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[17] 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.

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

[19] 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.

[20] I. Bica, A. M. Alaa, and M. van der Schaar, "Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders," in *Proceedings of the 37th International Conference on Machine Learning*, JMLR.org, 2019.

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

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

[23] 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.

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

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

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

[27] 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.

[28] 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.

[29] 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.

[30] J. R. Gimenez, A. Ghorbani, and J. Zou, "Knockoffs for the mass: New feature importance statistics with false discovery guarantees," in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* (K. Chaudhuri and M. Sugiyama, eds.), vol. 89 of *Proceedings of Machine Learning Research*, pp. 2125–2133, PMLR, 2019.

[31] A. Jesson, P. Manshausen, A. Douglas, *et al.*, "Using non-linear causal models to study aerosol-cloud interactions in the southeast pacific," in *NeurIPS 2021 Workshop on Tackling Climate Change with Machine Learning*, 2021.

[32] 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.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference for Learning Representations (ICLR)*, 2015.

A Appendix

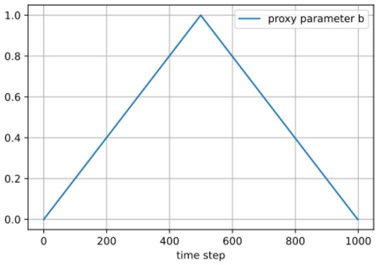

Figure 4: Time-varying proxy parameter b_t over 1000 time steps.

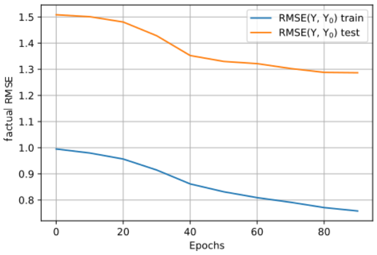

Figure 5: RMSE during training (blue) and test (orange) for the synthetic data with nonlinear causal link and time-varying proxy parameter b.

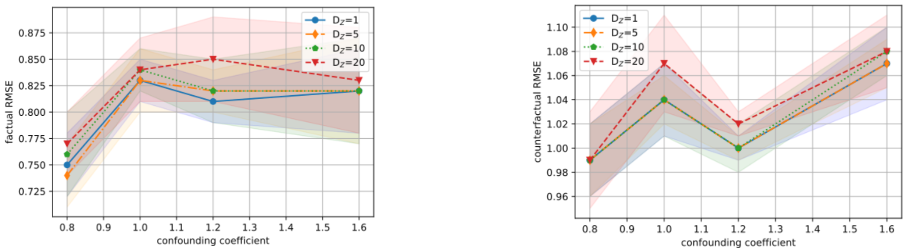

Figure 6: Influence of latent dimension and coefficient of the latent confounder to the outcome forecast with standard normal intervention. Factual RMSE is shown on the left and counterfactual RMSE on the right with standard deviation after five replications.