Knowledge and Information Systems (2021) 63:3041–3085
https://doi.org/10.1007/s10115-021-01621-0

SURVEY PAPER

Causal inference for time series analysis: problems, methods and evaluation

Raha Moraffah[1] · Paras Sheth[1] · Mansooreh Karami[1] · Anchit Bhattacharya[1] · Qianru Wang[2] · Anique Tahir[1] · Adrienne Raglin[3] · Huan Liu[1]

Received: 5 February 2021 / Revised: 21 October 2021 / Accepted: 23 October 2021 /
Published online: 23 November 2021

Abstract
Time series data are a collection of chronological observations which are generated by several domains such as medical and financial fields. Over the years, different tasks such as classification, forecasting and clustering have been proposed to analyze this type of data. Time series data have been also used to study the effect of interventions overtime. Moreover, in many fields of science, learning the causal structure of dynamic systems and time series data is considered an interesting task which plays an important role in scientific discoveries. Estimating the effect of an intervention and identifying the causal relations from the data can be performed via causal inference. Existing surveys on time series discuss traditional tasks such as classification and forecasting or explain the details of the approaches proposed to solve a specific task. In this paper, we focus on two causal inference tasks, i.e., treatment effect estimation and causal discovery for time series data and provide a comprehensive review

✉ Raha Moraffah
rmoraffa@asu.edu

Paras Sheth
psheth5@asu.edu

Mansooreh Karami
mkarami@asu.edu

Anchit Bhattacharya
abhatt22@asu.edu

Qianru Wang
qr369wang@gmail.com

Anique Tahir
artahir@asu.edu

Adrienne Raglin
adrienne.raglin2.civ@mail.mil

Huan Liu
huanliu@asu.edu

1 Computer Science & Engineering, Arizona State University, Tempe, AZ, USA

2 Northwestern Polytechnical University, Xi'an, China

3 Army Research Lab, Adelphi, USA

of the approaches in each task. Furthermore, we curate a list of commonly used evaluation metrics and datasets for each task and provide an in-depth insight. These metrics and datasets can serve as benchmark for research in the field.

Keywords Time series · Causal inference · Causal effect estimation · Causal discovery · Causal benchmarking · Causal evaluation · Granger causality · Structural causal models

1 Introduction

Time series data consist of ordered sequences of real-valued data which are often collected over time. With the rapid growth of time series data generated by different domains such as bioinformatics, medical, neuroscience and financial applications, various approaches have been developed. Research on time series data has been going on for over a decade, and researchers have come up with different approaches to analyze this type of data for different purposes such as classification [102,128], clustering [94,109], forecasting [76,169], estimating the impact of an intervention/treatment over time [14,116] and discovering the causal relations between the time series components [48,58]. In this survey, we focus on the last two tasks, i.e., estimating the effect of an intervention/treatment and identifying the causal relations and refer to them as causal inference for time series analysis.

Questions such as “Was an enforced policy effective?" or “Which medicine works better for a specific disease?" are crucial questions in law-making and medical fields, and answers to these questions can help making important decisions/policies. In order to answer such questions with data, one needs to estimate the effect of an intervention/treatment. For example, to evaluate the effectiveness of tobacco control program, Abadie et al. [1] propose a framework to estimate the effect of this program on cigarette sales. In another example, Bica et al. [26] propose a framework to predict patient’s response to a specific medicine over time. The task of estimating such effects from the data is called causal treatment effect estimation. Causal treatment effect estimation is one the most important tasks in causal inference which leverages concepts from causality to estimate this effect from the data. The state-of-the-art causal treatment effect estimation approaches for time series data can be categorized into three main types: (1) time-invariant treatment effect; (2) time-varying treatment effect; and (3) dynamic regimes. We discuss some widely used methods in each category.

Another principal task discussed in this survey is causal discovery. Causal discovery is the task of identifying the causal relationships between variables in the data. Causal discovery for time series data refers to the task of understanding and identifying interdependencies among individual components of a time series. This task is seen in a variety of applications such as economy and earth system science. For instance, causal discovery can be used to identify the performance indicators of stock analysis [70] or discover the causal relations between the external drivers of climate change and climate variables [141,158]. We classify causal discovery approaches for time series data into three main categories, namely, Granger causality and conditional independence-based, structural equation model-based and deep learning-based methods and discuss them in detail.

Despite having extensive surveys on non-causal time series analysis from different perspectives such as time series classification [3], deep learning and unsupervised feature learning [49] and data mining approaches [50], no existing survey reviews the current progress of algorithms which are designed to analyze time series data from a causal perspective. Different from existing efforts, in this paper, we discuss the state-of-the-art methods for causal

inference for time series analysis and its two main tasks. Since the evaluation of causal inference in general and causal inference on time series in particular is a challenging task, we also enlist some benchmark datasets and evaluation metrics which are commonly used by the researchers.

We first explain prevalent methods and concepts used for modeling time series data (Sect. 2). We then discuss necessary definitions and assumptions for causal inference on time series data which are used in the rest of the paper (Sect. 3). Next, we discuss causal treatment effect estimation and causal discovery for time series (Sects. 4.1 and 4.2, respectively). We then provide guidelines on how the frameworks proposed for each of these tasks can be evaluated by presenting a list of commonly used datasets and evaluation metrics (Sect. 5). We conclude this work with some future directions (Sect. 6).

2 Modeling time series data

Time series data are a sequence of real-valued data with each data point related to a timestamp. Mathematically, time series data are denoted as $\mathrm{X(t)} = (x_1(t), x_2(t), \ldots, x_k(t))$ where k is the number of variables measured at a discrete timestep $\mathrm{t} \in \mathbb{Z}$. In this section, we discuss different techniques to model a time series data.

2.1 Autoregressive models

One of the earlier methods to model time series data is the AutoRegressive Integrated Moving Average (ARIMA) model. The ARIMA models the time series assuming three fundamental relationships between the time series—autoregressive, moving average and differencing. The Autoregressive (AR) component determines the value of a current timestamp $X(t)$ from a finite set of previous timestamp values of some length p and some error ϵ. The order of autoregression is the number of preceding timestamps used to determine the value of the current timestamp, given by

$$AR(p) = \sum_{i=1}^{p} a_i . X(t-i) + c + \epsilon_t, \tag{1}$$

where a_i and p are the coefficients and the order of the AR model, respectively. The moving average (MA) component models the value of a current timestamp $X(t)$ as a linear combination of the prediction errors(ϵ_t) at the previous timestamps of length q, where q is the order of the moving average component.

$$MA(q) = \sum_{i=1}^{q} b_i . \epsilon_{t-i} + \mu + \epsilon_t, \tag{2}$$

where b_i, μ and q are the coefficients, mean of the series and the order of the MA model, respectively.

The AR and MA components are enough to model a time series in the case the time series data is stationary, i.e., the time series have the same values of specific properties (mean, variance) over every time interval. In the case of non-stationary data, as shown in Eq. 3, the time series data are differenced with a shifted version of itself to make the data stationary.

$$Y(t) = X(t) - X(t-r), \tag{3}$$

where r is the order of differencing.

The primary task in this type of modeling is estimating the coefficients a_i in AR and b_i in MA models, as well as the orders p, q, and r of the AR, MA and differencing, respectively. The Box and Jenkins [28] approach is used to estimate the parameters of an ARIMA model, assuming an underlying model, and verifying if the residuals or error term is a random distribution. This process is repeated with different models until the right model is obtained. The Simple Exponential Smoothing (SES) [32] is a modification of the ARIMA model with exponential weights assigned to each observation. Double and Triple Exponential Smoothing (DES and TES) [76] handles non-stationary data by introducing an additional parameter β for smoothing the trend in the series. Moreover, an extra parameter γ is introduced in TES to control the influence of seasonality.

2.2 Dynamic Bayesian networks

Dynamic Bayesian Networks (DBNs) are an extension of Bayesian Networks (BNs) to model the evolution of random variables as a function of a discrete timestep sequence, represented as a directed acyclic graph. Formally, a Bayesian Network is defined by $G = (V, E)$, where V and E are the set of nodes and edges. The conditional probability distribution of the set of nodes V can be expressed as the factorized joint probability given by:

$$P(V) = \prod_{x \in V} P(x|\pi_x), \tag{4}$$

where π_x are the parents of node x. A DBN is represented as a pair of two Bayesian networks B_p and B_{2d}. B_p is a BN modeling the prior distribution of the random variables at time 1. B_{2d} is a two slice BN representing the transition from time t−1 to time t, as a probability distribution P($x_t|x_{t-1}$) for nodes x belonging to V by means of a directed acyclic graph G = (V, E) as follows:

$$P(V_t|V_{t-1}) = \prod_{x \in V, \pi_x \in V} P(x_t|\pi_{x_t}) \tag{5}$$

If we define T as the total length of the path, the joint distribution of the sequence is given by:

$$P(V_{0:T}) = \prod_{x \in V} P_{B_p}(x_1|\pi_{x_1}) \times \prod_{t=2}^{T} \prod_{x \in V} P_{B_{2d}}(x_t|\pi(x_t)) \tag{6}$$

Typically, the variables in a DBN are partitioned into two sets of variables, $V_t = (Z_t, X_t)$, representing the hidden and output (observed) variables of a state-space model.

The parameters of the DBN can be learned from the data. Based on the probability distributions and the assumptions made on the dynamics (in the case of observable data), Maximum Likelihood Estimation (MLE) or Maximum A Priori (MAP) is used. For hidden variable models, the parameters are generally learned using the Expectation-Maximization (EM) algorithm.

Next, we discuss two of the state-of-the-art DBN models used for time series modeling.

2.2.1 State-space models

State-space models use a latent state z_t to model the time series data, i.e., encoding time series components level, trend and seasonality patterns. An SSM is denoted by a state-transition equation, which describes the transition dynamics $p(z_t|z_{t-1})$ of the evolution of

the latent state over time. It also represents an observation model that describes the conditional probability $p(x_t|z_t)$ of observations given the latent state. A widely used example of SSM is the linear dynamical system (LDS), where the states are real-valued and change linearly with time, satisfying the first-order Markov assumption. The LDS can be expressed by the following equations:

$$z_t = Az_{t-1} + \eta, z_t \in \mathbb{R}^k, \eta \sim N(0, Q) \tag{7}$$

$$x_t = Cz_t + \epsilon, x_t \in \mathbb{R}^d, \epsilon \sim N(0, R), \tag{8}$$

where $A \in \mathbb{R}^{kXk}$, $C \in \mathbb{R}^{dXk}$, while Q and R are covariance matrices. The joint probability for the states and observations is given by:

$$P(z_t, x_t) = P(z_1)P(x_1|z_1)\prod_{t=2}^{T} P(z_t|z_{t-1})P(x_t|z_t) \tag{9}$$

The Kalman filtering algorithm [85] is used to perform inference tasks such as filtering ($p(z_t|x_{1:t})$), smoothing ($p(z_t|x_{1:T})$), and prediction ($p(x_t|x_{1:t-1})$).

Nonlinear extensions of the LDS is proposed in unscented Kalman Filter [84], by generalizing the state transition and emission to nonlinear functions. Zheng et al. [183] propose State-space LSTM which modeled the sequential latent states by parameterizing the transition function between states by a neural network. The Particle Gibbs [7] is used for parameter estimation of the model which samples from the joint posterior, eliminating the need to sample at each time point thus removing the assumption of a factorizable posterior. Karl et al. [86] uses the stochastic gradient variational Bayes to learn the latent state dynamics under a nonlinear Markovian setting. Instead of a deterministic f and g functions, Wang et al. [168] placed Gaussian process priors over both the nonlinear functions f and g and found a MAP estimate of the latent variables.

Switching state-space models [53], model the observations x_t using M real-valued hidden state-space vectors z_t^m and one discrete state vector s_t. A multinomial variable with M possible values represents the discrete state variable $s_t \in (1, \ldots, M)$ which acts as a switch variable. The observed variable is represented using the state-space model m, conditioned on this discrete state. The discrete state follows a Markovian dynamics with a specified initial state($p(s_1)$) and transition probability matrix($p_{s_t|s_{t-1}}$). The real-valued state variables have linear-Gaussian dynamics with each variable having its transition matrix, initial state and noise. The joint distribution of the observed and hidden variables is given by:

$$\begin{aligned} P(s_t, z_t^1, \ldots, z_t^m, x_t) = P(s_1)\prod_{t=2}^{T} P(s_t|s_{t-1}) \\ \times \prod_{m=1}^{M} P(z_1^m)\prod_{t=2}^{T} P(z_t^m|z_{t-1}^m) \\ \times \prod_{t=1}^{T} P(x_t|z_t^1, \ldots, z_t^M, s_t) \end{aligned} \tag{10}$$

The discrete switch variable acts as a gating network for the M real-valued states. Gibbs sampling is used to approximate the marginal probabilities, required for the evaluation of the expectations for learning the model parameters using the EM algorithm.

2.2.2 Hidden Markov model

The Hidden Markov Model (HMM) is another class of SSM, where the states are assumed to be discrete and distributed according to the Markov Process. The joint probability distribution of the discrete hidden state s_t and the observed sequence x_t can be denoted similarly to the SSM as:

$$P(s_t, x_t) = P(s_1)P(x_1|s_1)\prod_{t=2}^{T} P(s_t|s_{t-1})P(x_t|s_t) \tag{11}$$

The Baum–Welch algorithm [21], a type of EM algorithms for HMM, is used to learn the parameters of this model.

In many applications, the HMM model is extended by representing the $P(x_t|s_t)$ using a mixture of Gaussians for each state (M_t), forcing x_t to get the information from s_t bottlenecked through M_t.

Autoregressive HMMs [149] relaxes the HMM assumption of conditional independence of observations given the hidden state, by allowing x_t to be connected to x_{t-1} along with the hidden state s_t.

Factorial HMMs [54] extend the HMM by having a collection of discrete state variables, as opposed to a single state variable for the original HMM. The state variable s_t is represented as a combination of $(s_t^1, \ldots, s_t^M)$, each of which can take K possible values. This will extend the state space to K^M possible values, which is equivalent to a regular HMM with K^M states. Ghahramani et al. [52] propose a constraint on the interactions of the state variables, in which the state variables evolve through the following dynamics:

$$P(s_t|s_{t-1}) = \prod_{m=1}^{M} P(s_t^m|s_t^{m-1}), \tag{12}$$

thus uncoupling the different state variables from each other.

The uncoupling of state variables was relaxed in Saul et al. [145] by coupling the variables of a time step in order, i.e., s_t^m interacts with s_t^n for $1 \leq \mathrm{n} < \mathrm{m}$. The model parameters are learned using the EM algorithm, where the marginal probabilities used for the expectations are approximated using the Gibbs sampling algorithm.

2.3 Gaussian processes

Taking advantage of available data and performing robust analysis is only practical through modeling the uncertainty. Therefore, the Bayesian inference is leveraged to handle uncertainty in a noisy and dynamic environment. Gaussian Processes are a class of Bayesian nonparametric models that are particularly suitable for modeling time series data. In particular, Gaussian Processes (GPs) are a class of stochastic processes, which define a joint Gaussian distribution over a collection of random variables. A function ($f(x)$) which follows a Gaussian process is specified by the mean (m(x)) and covariance ($k(x, x^{'})$ functions, denoted as $f(x) \sim GP(m(x), k(x, x^{'}))$. Formally, the Gaussian process responsible for generating Y given X is given by,

$$\mathbf{y}_n = f(\mathbf{x}_n) + \epsilon_n, \quad \epsilon \sim \mathcal{N}\left(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}\right) \tag{13}$$

where ϵ is the Gaussian noise term. In what follows, we briefly discuss the Gaussian processes frameworks.

2.3.1 Deep learning and GP

Motivated by the success of Deep models in different tasks, recently, several attempts have been made to combine deep models with Gaussian processes [43,110,170,171] and create Deep Gaussian Process (Deep GP) models. These frameworks typically use neural networks to map the input to the feature space (extract non-stationary features), whereas the last layer sparse Gaussian process performs regression over the latent space. For example, Wilson et al. [170] propose to leverage fully connected and convolutional neural networks as input to the spectral mixture base kernel and use local kernel interpolation [173], spectral mixture covariance functions [172], inducing points [126] and structure exploiting algebra [142] and create more powerful and expressive closed-form covariance kernel for Gaussian Processes. In another attempt, Maddix et al. [110] propose a scalable hybrid model which combines both deep neural network and classic time series model to perform accurate forecasting which also takes uncertainty into account. The model consists of a global deep neural network and a local Gaussian Process model.

2.3.2 GP methods for inference

Havasi et al. [65] propose an inference method for Deep Gaussian Processes models based on the Stochastic Gradient Hamiltonian Monte Carlo method. The paper also shows that the posterior in these models is of non-Gaussian nature, and therefore, the existing approaches based on the variational inference that estimate a Gaussian posterior are poor potential approximations for the multimodal posterior.

The authors in [97] propose an uncertainty-aware classification framework that facilitates learning black-box classification models for classifying sparse and irregularly sampled time series. The framework uses Gaussian Process regression to transform the irregular time series data into a uniform representation, which permits sparse and irregularly sampled data to be fed into any black-box classifier which is learnable using gradient descent while preserving uncertainty. Tobar et al. [165] propose a framework called Gaussian Process Convolutional Model (GPCM), which serves as a generative model for stationary time series. The main idea behind this model is based on the convolution between a filter function and a white noise process. This approach recovers a posterior distribution over the spectral density directly from the time series. It also places the nonparametric prior over the spectral density before recovering the posterior distribution. Learning the model from the data allows performing inference on the covariance kernel as well as the spectrum in a probabilistic, analytic and computationally tractable manner. Cunningham et al. [42] propose a Gaussian Process model for analyzing multiple time series with multiple time markings. The proposed model can be considered as a mapping between the input space and the given data markers. Because of this, the model can be used as a choice for a covariance function. It also facilitates learning and inference to be standard. HajiGhassemi and Deisenroth [62] propose an algorithm for long-term forecasting with periodic Gaussian processes. They also state that, for long-term forecasting it is necessary to map the probability distributions through Gaussian Processes. They use re-parameterization of a commonly used stationary periodic kernel which in turn allows them to employ an analytic double approximation strategy to compute the moments of the predictive distribution.

For a comprehensive surveys on Gaussian Processes and how they can be used to model time series data and scaled to big data scenario, we refer the reader to three studies provided by Rasmussen [127], Roberts et al. [129], and Liu et al. [104].

2.4 Neural networks

For more complex, noisy, and higher dimensional real-world data, techniques such as ARIMA and state-space models are not efficient since the dynamics are unknown or too complex [163]. Various unsupervised deep models have been extended for time series data to solve this problem. In the following, we discuss some of the state-of-the-art deep learning models proposed for time series. For more details, we refer the readers to the recent surveys on deep learning for time series [49,51,100].

Graves et al. [59] propose a method for sequence generation using recurrent models such as RNN and LSTM, by processing the real data at each step (x_t) and predicting the value of the next step (x_{t+1}). Output predictions (y_t) at each step are made probabilistic and sampled from to be fed as the next step input. Iteratively sampling at each step from the already trained network, and passing it to the next step produces a novel sequence. Although theoretically possible, RNNs in practice suffer to capture long-term dependencies, and thus, LSTMs are used which are shown to capture long-term dependencies with the help of various gating mechanisms. Additionally, the hidden layers are stacked to increase the depth across space to allow for higher nonlinearities to be captured at each time step. The probability of the input sequence x is given by $\Pr(\mathrm{x}) = \prod_{t=1}^{T} Pr(x_{t+1}|y_t)$.

Restricted Boltzmann Machine(RBM), a generative probabilistic model between the input nodes (observable) and latent nodes (hidden), connected by a weight matrix (W) and having associated bias vectors c and b, respectively, is extended for sequential data in various works. The model is generally trained by minimizing the reconstruction error using contrastive divergence [71]. Conditional RBM and temporal RBM [161] extend the RBM model with a connection between the current hidden units and the past observable units along with auto-regressive weights for capturing short-term temporal patterns. The dependency between the bias vectors and the past visible units are defined by,

$$\begin{aligned} b_j^{'} &= b_j + \sum_{i}^{n} B_i x(t-i) \\ c_i^{'} &= c_i + \sum_{i}^{n} A_i x(t-i), \end{aligned} \tag{14}$$

where B_i and A_i are the weight matrices connecting, respectively, current hidden units and current observable units to observable units at time $t-i$. Thus, the conditional probabilities for activation of the hidden units and the visible units become:

$$\begin{aligned} P(h_j|x) &= \sigma\left(b_j + \sum_i W_{ij} x_i\right) + \sum_k \sum_i B_{ijk} x_i (t-k) \\ P(x_i|h) &= \sigma\left(c_i + \sum_i W_{ij} h_j\right) + \sum_k \sum_i A_{ijk} x_i (t-k) \end{aligned} \tag{15}$$

Oord et al. [119] propose a CNN-based architecture called WaveNet to generate audio waveforms. WaveNet tries to approximate the joint probability of the time series $X = (x_1, x_2, \ldots, x_T)$ by making x depend on all the previous samples. WaveNet uses a special convolutional layer called the dilated causal convolution. A causal convolution forces the prediction at any timestep to depend on the previous timesteps and prevents dependencies on future timesteps. Dilated convolutions is a filter which can span larger than it's length by dilation with zeros, skipping input values at some steps, thus increasing the receptive field of

the filters with increasing depth. Additionally, gated activated units are used which allows the network the ability to preserve and forget certain input values. Undecimated Fully Convolutional Network (UFCNN) [113] uses Fully Convolutional Layers with 1D causal filters, and the filter's at lth resolution level is upsampled by a factor of 2^{l-1} along the time dimension along with removal of max-pooling layers and other upsampling operators.

Generative Adversarial Network (GAN)-based models have also been proposed for Time Series Generation. Mogren et al. [115] propose C-RNN-GAN to generate continuous sequential data by modeling the joint probability distribution of the sequence. The generator was designed using an LSTM, and the discriminator consisted of a bidirectional RNN. The model was trained using the standard GAN loss. Yoon et al. [181] argued that using recurrent networks for the generator and discriminator, and summing the GAN loss over sequences is not enough for capturing the temporal dynamics of the data. They propose a stepwise supervised loss along with the unsupervised adversarial loss to encourage the model to capture the stepwise temporal dependencies.

3 Preliminaries and background

In this section, we briefly discuss the definitions, notations and assumptions from causal inference for time series literature which are used throughout the paper. We start with introducing common definitions and assumptions in causal treatment effect estimation. In order to calculate the difference between the outcome of the data with intervention and the control group, several metrics have been utilized. In the following, we define the common metrics and review the required assumptions to make these estimators consistent.

Suppose A is a treatment dummy random variable and Y is the desired outcome random variable, then A can take value $a = 1$ or $a = 0$ denoting the presence or absence of a treatment, respectively. In this case, Y_a will be defined as the potential outcome under exposure to the treatment (i.e., $Y_{a=1}$) or under control (i.e., $Y_{a=0}$). $\delta Y_a = Y_{a=1} - Y_{a=0}$ denote the individual (or unit) causal effect (ITE) of the treatment.

Definition 1 (Average Treatment Effect) The average causal effect also known as the average treatment effect of the population can be calculated as follows:

$$ATE = \mathbb{E}[Y_{a=1} - Y_{a=0}], \tag{16}$$

and it is nonzero when the treatment A has a causal effect on the mean of the outcome.

This measure implicitly assumes that the individuals are drawn from a large population. However, due to selection bias, the units might not be representative of such a population. In this case, sample average treatment effect (SATE) will be used that only calculates the treatment effect of the units in that specific study, which avoids any assumption on the distribution of the samples [19].

Definition 2 (Sample Average Treatment Effect)

$$SATE = \sum_{i=1}^{m} [Y_{i,a=1} - Y_{i,a=0}], \tag{17}$$

where m is the number of samples and $Y_{i,a}$ is the outcome of sample i under the treatment a.

On the one hand, individual treatment effects in the population might be heterogeneous, meaning that the treatment affects the individuals or sub-populations differently. In this

case, it is more desirable to consider the conditional average treatment effect (CATE) [4] to calculate the effect of a treatment on the sub-population.

Definition 3 (Conditional Average Treatment Effect) In the randomized controlled trial, the Conditional Average Treatment Effect (CATE) is estimated as follows:

$$CATE = \mathbb{E}[Y_{a=1} - Y_{a=0}|X = x], \tag{18}$$

where X is the covariates (or features) and x is the values that the covariates take.

On the other hand, we might be interested in the causal treatment effects for only those individuals of the population who choose to participate in the treatment. In this case, we can calculate the average treatment effect of the treated (ATT) sub-population as follows:

Definition 4 (Average Treatment Effect of the Treated)

$$ATT = \mathbb{E}[Y_{a=1} - Y_{a=0}|A = 1] \tag{19}$$

All the above estimators are true if we conduct experiments on randomized trials. However, researchers may only have access to outcome values reported at the aggregate level (observational data). To obtain consistent estimators from observational data, identifiability conditions should be hold [67,68,136]:

Assumption 1 (Consistency) This assumption indicates that if $Y_{a=1}$ denotes the potential outcome for the treated subject, then its value is known and is equal to the observed outcome, Y. Although $Y_{a=0}$ (potential outcome under control) remains unknown. This is also true for the untreated subject. In other words, for the treatment variable A, if $A = a$, then $Y_a = Y$.

Assumption 2 (Positivity) The probability of receiving every value of treatment conditional on some measured covariates X is greater than zero. In other words, $Pr[A = a|X = x] > 0$ for all values x with $Pr[X = x] \neq 0$, in the population of interest.

Assumption 3 (Conditional Exchangeability) Unconditional exchangeability implies that the treatment group, had they been untreated, would have experienced the same distribution of outcomes as the control group. While in conditional exchangeability, the conditional probability of receiving every value of treatment, depends only on measured covariates X, i.e., Y_a and A are statistically independent given every possible value for X.

In order to be able to perform causal discovery, additional assumptions need to be made. Below we list some common assumptions required to perform causal discovery for time series data.

Assumption 4 (Causal Stationarity) The time series process X with graph defined over it is called causally stationary over a time index set T if and only if for all links x^i_{t-T}

$$x^i_{t-T} \not\perp\!\!\!\perp x^j_t | X^-_t \backslash x^i_{t-T} \text{ holds for all } t \in T. \tag{20}$$

Assumption 5 (Causal Sufficiency) A set of variables is causally sufficient for a process, if and only if it includes all common causes of every two pairs in the set.

Assumption 6 (Causal Markov Condition) The joint distribution of a time series process X with graph G fulfills the Causal Markov Condition if and only if for all $Y_t \in X_t$ with parents P_{Y_t} in the graph:

$$X^-_t \backslash P_{Y_t} \text{ d-separated } Y_t | P_{Y_t} \implies X^-_t \backslash P_{Y_t} \perp\!\!\!\perp Y_t | P_{Y_t}. \tag{21}$$

Assumption 7 (Faithfulness) The joint distribution of a time series process X with graph G fulfills the Faithfulness condition if and only if for all disjoint subsets of nodes (or single nodes) $A, B, S \subset G$ it holds that

$$X_A \perp\!\!\!\perp X_B | X_S \implies A \text{ d-separated } B|S. \tag{22}$$

4 Causality and time series analysis

In this section, we discuss two most important causal inference tasks for time series data, i.e., causal treatment effect estimation and causal discovery.

Causal effect estimation refers to the task of estimating the effect of a policy or treatment on a target variable. This effect is commonly measured with metrics such as the Individual Treatment Effect (ITE), Average Treatment Effect (ATE), Conditional Average Treatment Effect (CATE) and Average Treatment Effect on the Treated (ATT). These metrics are explained in Sect. 3.

Causal Discovery is the task of identifying causal relationships between variables in the system from the data.

In the following, we first explain causal treatment effect estimation problem, classify the existing approaches based on different settings given the time and explain the state-of-the-art approaches in each category. We then discuss the task of causal discovery for time series data, categorize the proposed frameworks based on the type of model being used and explain them in detail. Tables 1 and 2 illustrate compiled lists of proposed methods for each task along with their applications. Note that these applications are the ones used in the original paper's experiments.

4.1 Causal treatment effect estimation on time series

Policymakers often face challenges to assess the impact of an intervention (i.e., a change in policy) on an outcome of interest. For example, a state government wants to estimate the effect of a tobacco control program on cigarette sales using the available data before and after a proposition [1].

There is a need to evaluate both the positive and negative valued consequences of the designed or unintended policies and interventions to ascertain whether they were effective or not.

To this end, researchers proposed various methodologies that account for different settings based on the time to estimate the treatment effect: (1) time-invariant treatment effect, (2) time-varying treatment effect, and (3) dynamic regimes.

In this section, we will introduce the recent developments and existing applications of causal treatment effect estimation based on the aforementioned categories specifically designed for time series data.

4.1.1 Time-invariant treatment effect

A treatment is time-invariant or fixed when it occurs at one specific point of time and then does not change afterward, for example, a one-dose drug. More formally, let $X(t)$ be the time series outcome recorded at times $t = 1, 2, \ldots, n$, and let A be a dichotomous treatment that can take values $a = 0$ (untreated) or $a = 1$ (treated). We will have X_a that denotes the

Table 1 Overview of causal treatment effect estimation methods and their applications

Category	Method	Papers	Type of time series	Characteristics/advantages	Example applications
Time-invariant treatment effect	Difference-in-differences (DiD)	[8,9,12,13,174]	Linear/nonlinear time series	This method is appropriate when the outcome variables under study are *not serially correlated* and follows the *common trend assumption*.	Various (economics, sociology, healthcare, marketing)
	Causal impact	[30]	Bayesian structural time series	Estimates the counterfactual of the time series of only *a single treated unit* based on diffusion-regression state-space model even if the time series data is highly *auto-correlated*	Incremental impact of market interventions, e.g., the causal effect of an online advertising campaign on search-related site visits
	Causal transfer	[95,96]	Linear/Nonlinear non-Stationary time series	An algorithm using linear state-space model for *heterogeneous treatment effect estimation* in time series. This method can be applied to *non-stationary* randomized trials and observational studies in which the treatment is *confounded*. It also provides information on *long-term effects*	Various estimands in both forms of the population or sample version
	Synthetic control method (SCM)	[1,2,16,35,39,91, 114,146]	Linear time series	This method is used to estimate the effect of an intervention on *aggregate/macro level data* when there is an *ambiguity* on how to choose the control units. The number of control units should be small	Economics, health policies, or crime interventions in countries, regions, or cities level
		[5]	Linear/nonlinear time series	This method is used when there is a *missing data* or there is a *large level of noise* (A robust SCM). This algorithm also performs well *in the absence of covariate or expert information*	Economics, health policies, or crime interventions in aggregate level
	Interrupted time series (ITS)	[23]	Linear/Nonlinear time series	This method is appropriate for cases when the *outcome lasts long enough* to measure and *changes relatively fast* after the intervention or after a defined lag Moreover, we should also be able to place the *exact time of the intervention*. The standard ITS is used when the *control unit data is not available*	Healthcare
		[101]	Linear/Nonlinear time series	This version can be used when *one or more control groups* are also available for comparison	Economics and health policies

Table 1 continued

Category	Method	Papers	Type of time series	Characteristics/advantages	Example applications
Time-varying treatment effect	Structural nested models (SNM)	[130,131]	Linear/nonlinear time series	Estimates the causal effect in the presence of *time-varying covariates*	Various (mostly healthcare and epidemiology)
	Marginal structural modeling (MSM)	[69,134]	Linear/Nonlinear time series	Estimates the causal effect in the presence of *time-varying covariates*. This method is an alternative to SNM. MSM requires *more parametric assumptions*, but works better than SNM in the *absence of time-dependent confounders*	Various (mostly healthcare)
	Bayesian nonParametric (BNP)	[153,179]	Bayesian multivariate time series	Estimates the trajectories of the treatment response from *sparse observational time series*	Various (mostly healthcare)
	Deconfounder	[26]	Multiple treatment time series	Estimates the individual response to treatments in the presence of *multi-cause hidden confounders* using factor models	Healthcare
	Deep sequential weighting (DSW)	[105]	Linear/nonlinear dynamic time series	Estimates the individual response to treatments with *time-varying confounders* when there also exists *hidden confounders*	Economics and health policies
Dynamic Treatment Regimes	Regret Function	[116]	Complex time series data	This method can be used when you are looking on how treatment should be allocated overtime. It also provides formulae for transforming between blip and regret functions	Healthcare
		[118,132]	Discrete time series data	Provides *optimal* dynamic treatment regimes (maximal mean response). This method can be used when in any time interval, *treatment decisions are made on all or most subjects*. In other words, the timings of the treatment decisions should not be variable across subjects	Healthcare
	Lagged treatment effect	[184]	Continuous treatment	This method provides dose suggestions which *maximize short-term outcomes*. Contrary to other methods where stationarity or Markovian property is required, this method imposes *minimal assumptions* on the data generating process	Healthcare

Table 2 Overview of causal discovery methods and their applications

Method	Paper	Type of time series	Characteristics/advantages	Example applications
Granger causality based models	[60]	Linear, stationary time series	This method incorporates possible *shared structures* between time series in the causal graph into forecasting	Real-life forecasting
	[151]	Linear time series	This method is suitable for the case where #components >> #samples. It determines order of VARs and reduces the number of covariates	Estimation of gene regulatory networks
	[152]	Linear/nonlinear, low and high dimensional time series	This method is suitable for the case where #components >> #samples by determining the order of VARs	EEG data analysis
	[6]	Linear time series	This method is used when *the lags between cause and effect have variable length*	Studying coordinated collective behavior
	[63]	Multivariate linear time series	Learns *sparse causal graph* by zeroing out all coefficients of AR for pairs of time series with no causal relations	Synthetic data analysis
Conditional independence based models	[159,160]	Nonlinear time series	This method infers the causal relations *in the presences of indirect connections, dominance of neighboring dynamics, or anticipatory couplings.*	Studying the dynamics of small scale coupled oscillators networks
	[48]	Multivariate time series	This approach takes *latent confounding variables* into account.	Analysis of microeconomic data of growth rates of US manufacturing
	[138]	Linear/Nonlinear Time series	This approach is specifically designed for *high-dimensional* and *nonlinear data with limited sample sizes*	Climate teleconnection
	[140]	Linear/Nonlinear time series	Discovers *contemporaneous links* along with lagged links	Synthetic data analysis
	[38]	Nonlinear, stationary time series	Develops *conditional independence test for nonlinear systems*	Studying the ocean climate indices
	[82]	Nonlinear time series	Reports Leading time in time series, Leverages Pearl's causality	Hydrological study

Table 2 continued

Method	Paper	Type of time series	Characteristics/advantages	Example applications
Structural equation based models	[78]	Linear non-Gaussian time series	Extends LiNGAM to *combine Instantaneous and Lagged Effects*	Finance and magnetoencephalography
	[147]	High dimensional linear non-Gaussian tensor data	Recovers causal relations for *high dimensional data*	Various applications such as medical domain
	[137]	Linear non-Gaussian	Takes *latent variables* into account.	Flow cytometry data analysis
	[73]	Nonstationary time series	Leverages a particular type of state space model to do causal discovery for *nonstationary time series*	Economics
	[123]	Independent residual time series	Avoids incorrect answers when *the data is causally insufficient* or *the model is misspecified*	Various including finance and economics
Deep learning based models	[162]	Nonlinear stationary time series	Proposes a class of *nonlinear methods* to identify causal relationships	Gene expression data analysis
	[44]	Linear/Nonlinear time series	Captures *nonlinear* and *time variant* dependencies	Cloud service data analysis
	[176]	Nonlinear time series	Discovering *nonlinear* causal relationships with *causal sufficiency* and *causality in mean* assumptions	Neural activity and locomotion data analysis
	[178]	Nonlinear time series	This method discovers *nonlinear* interactions and is specifically designed to scale to *high dimensional data*	Engine operating data analysis
	[120]	Multiple time series data corresponding to different spatial locations	Identifies causal relations by leveraging both *spatial* and *temporal* information	Air quality/ Traffic/ Flow prediction
	[112]	Nonlinear time series	Takes *unobserved confounders* into account	Economics

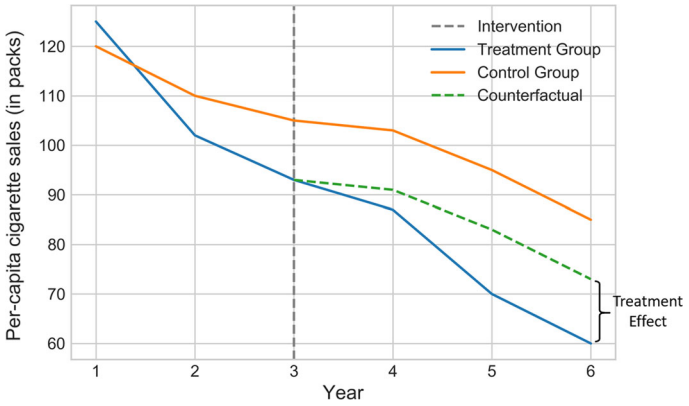


Fig. 1 An illustration of the DiD method on a hypothetical example. The difference between the green dotted line and the blue line after the intervention is the treatment effect

potential outcome for the treatment group (when $a = 1$) or control group (when $a = 0$). At $1 \leq T \leq n$ an intervention occurs. We are mostly seeking to model the potential outcome of the treatment group at $t > T$ had it not received the treatment ($X^c_{a=1}(t > T)$) which is known as the counterfactual outcome. The counterfactual tells us what would have happened in the treatment group had we not applied the policy. The difference between the observed values for the treatment group ($X_{a=1}(t > T)$) and the counterfactual outcome (unobserved values) would give us an estimation for the treatment effect. This difference can be reported as a numerical value using measures introduced in Sect. 3 such as ITE, ATE and ATT. In the following section, we will introduce common time-invariant causal effect estimation methods.

An important tool mostly used by the econometricians to capture the causal effect of the time series data before and after the treatment is *difference-in-differences* (DiD) [14]. This tool goes by an assumption called *common trends* or *parallel trends* [9] that uses the change of the outcome of the control group as a counterfactual for the treatment group in the absence of the treatment. Figure 1 illustrates a hypothetical example of DiD on cigarette sales before and after the tobacco control program. The aim is to see whether this program affected per-capita cigarette sales.

If $t < T$ and $t > T$ denote the pre- and post-treatment periods (e.g., $T = 3$ in Fig. 1), respectively, then we can calculate the DiD measure using the ATT metric as follows [13]:

$$\begin{aligned} DiD = &\{\mathbb{E}[X(t > T)|A = 1] - \mathbb{E}[X(t < T)|A = 1]\} \\ &- \{\mathbb{E}[X(t > T)|A = 0] - \mathbb{E}[X(t < T)|A = 0]\} \end{aligned} \tag{23}$$

A common solution to model the DiD causal effect is to specify a linear regression model for the observed outcome [9]. In case of one intervention and one treatment group, we will have:

$$X = \alpha + \beta g + \gamma t + \delta(g \times t) + e, \tag{24}$$

where α, β, γ and δ are the regression coefficients. g is a dummy variable indicating the treatment group (1 for treatment and 0 for the control group), t is a dummy variable defining the pre- or post-treatment periods (0 for before and 1 for after the treatment time), $g \times t$ is the interaction term to count the DiD causal effect with δ being its coefficient, and e is the error term which is independent of other variables. The parameters of this linear regression can be

estimated by the available methods solving regression models such as ordinary least squares (OLS) or gradient descent. Other variations of DiD regression consider settings with various treatments and multiple treatment groups [8,9], nonlinearity assumption for DiD model [13], and two control groups (known as triple differences) [12,174]. For a comprehensive survey on this metric, we refer the readers to two studies provided by Lencher [92] and Bertrand et al. [25].

Although in the design of the DiD, a temporal component is used, it has been pointed out that if it fits to highly auto-correlated data, the model underestimates the effect of the intervention [25]. To overcome this problem, Brodersen et al. [30] propose a method named *Causal Impact* which is widely used for various applications such as the impact of vaccines, the environmental impact of aircraft emissions and aviation fuel tax, and the impact of mobile phone use on brain cancer [33,57,166,167]. This method generalizes the concept of the DiD and structural time series model to infer the causal impact of a discrete intervention (e.g., the release of a new product). Causal impact learns the relationship between the treatment and control group before any intervention and predicts the counterfactual series after the treatment. This method relies on state-space models as follows [144]:

$$\begin{aligned} X(t) &= \mu(t) + \beta Z(t) + v(t) \\ \mu(t) &= \mu(t-1) + \delta(t-1) + w(t) \\ \delta(t) &= \delta(t-1) + u(t), \end{aligned} \tag{25}$$

where $Z(t)$ is the control time series and is related to the treatment time series (i.e., $X(t)$) through the β components. $v(t)$, $w(t)$ and $u(t)$ are zero-mean noise variables and $\mu(t)$ models the temporal correlation in $X(t)$. The $\delta(t)$ component can be thought of as the slope at time $t-1$ or the expected increase in μ between times $t-1$ and t. The model is fitted to the observed data $t = 1, 2, \ldots, T$, treating the counterfactual $t = T+1, T+2, \ldots, n$ as unobserved random variables. With these, the model will compute the posterior distribution of the counterfactual time series. The causal effect is estimated by subtracting the predicted from the observed treated time series, which captures the semi-parametric Bayesian posterior distribution. Li and Bühlmann [95,96] proposed a state-space-based model named *Causal Transfer*, inspired by *Causal Impact*, to learn the effect of the treatment in a time series data and capture how it evolves over time in order to transfer it to other time series. Specifically, after the treatment, we are only able to observe the outcomes under the treatment for one time series and under the control for another one, but not the potential outcome under control for the former and under treatment for the latter. The authors fill the missing outcomes by learning the intervention effect through a state-space model.

Synthetic Control Method (SCM) introduced by Abadie and Gareazabal [2] in 2003 overcomes the problem of ambiguities in the selection of control groups with the aim to estimate the effects of interventions that take place on an aggregate level (such as countries, regions, cities). In this method, we find the weight for each control unit such that the weighted average of all these potential control units (named as *donor pool*) best resembles the characteristics of the treated unit before the treatment and use the learned weights to estimate the counterfactual after the intervention. Formally, the SCM finds the weights by minimizing:

$$\begin{aligned} W^* &= \arg\min_W ||X_{a=1}(t < T) - X_{a=0}(t < T)W|| \\ &\text{s.t.} \sum_{i=1}^{J} w_i = 1 \text{ and } w_i \geq 1, \end{aligned} \tag{26}$$

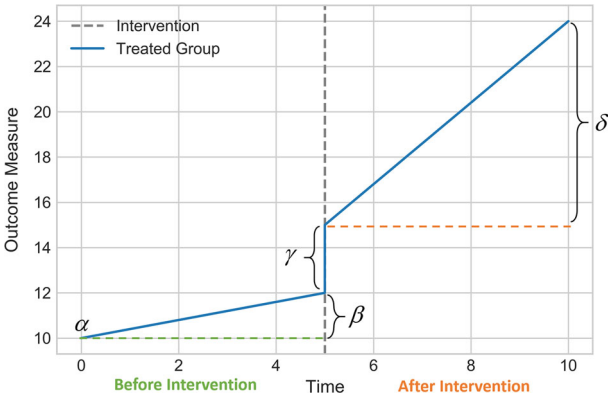


Fig. 2 An illustration of the Standard ITS method

where $X_{a=1}(t < T)$ is a $T \times 1$ treated unit vector and $X_{a=0}(t < T)$ is a $T \times J$ matrix with J being the number of control units. The predicted counterfactual of the treated unit is calculated by

$$X^c_{a=1}(t > T) = X_{a=0}(t > T)W^* \tag{27}$$

Other variations such as allowing for multiple treated units [91], estimation in the presence of missing data [5] and applications of this model [1,16,35,39,114,146] have also been conducted by researchers.

In the case when (1) the intervention begins at a known point in time, (2) the outcome changes relatively faster after the intervention or a defined lag, and (3) the outcome lasts long enough to measure [23], a method name *Interrupted Time Series* (ITS) analysis can be used. Although different variations of ITS exist [101], the standard ITS model is capable of finding the causal effect of an intervention for only one time series (i.e., when the control unit data is not available). This method is built upon a simple idea that the data generating process would have continued in a similar way in the absence of the intervention, which is a special case of *Regression Discontinuity Design* (RDD) when the discontinuity happens in time [64,90]. Therefore, to find immediate changes in the outcome value and the change in the trend of the time series in the post-intervention period compared to the pre-intervention period, "Segmented Regression" is used [122]:

$$Y = \alpha + \beta T + \gamma t + \delta(T \times t) + e, \tag{28}$$

where α indicates the baseline level at $T = 0$, β represents the baseline trend (i.e., pre-intervention trend), γ is the immediate level change following the intervention, and δ indicates the post-intervention trend and e is the error (Fig. 2).

Eichler and Didelez [46] proposed a formal definition of causality along with an identifiability criteria for estimating the causal effects based on the intervention in dynamic Bayesian networks. The authors considered the effect of intervention in one component of the multivariate time series, at a specific point of the time, on another or the same component at later time points. They assumed that the causal effect excludes instantaneous causality and depends only on past variables. Using graphical models for the time series data, they proposed that if (a, b) is not a directed edge in the graph, then the time series data of component a at time t have no causal effect on time series data of component b at time $t + 1$. Moreover,

if a is not the ancestor of b in the graph, then the time series data of component a at time t have no causal effect of the component b in the future.

4.1.2 Time-varying treatment effect

Most treatment effect estimation problems do not fit into the simple dichotomous treatment framework and require multiple sequential treatments which varies according to the time of the treatment [40]. For example, a drug dose when the dose is readjusted according to the patient's clinical response [135]. In this case, A, the treatment variable, would be time-dependent and is recorded throughout all the time ($A(t) = \{A(1), A(2), ...A(n)\}$). Below, we will briefly review the works in this area.

Works conducted by Robins et al. [133] as well as Hedeker and Gibbons [67] expand the definitions of consistency, positivity and conditional explainability assumptions along with the identifiability conditions of the time-invariant treatment effect estimation to estimate the causal effect of time-varying treatments. Moreover, in the presence of time-dependent confounding, counterfactual inference of a time-dependent treatment has been extensively studied in the epidemiology literature specially in the chain of works conducted by Robin [69,130,131,134] on *structural nested models* (SNM) and *marginal structural modeling* (MSM). MSM, which is an alternative to SNM [134], models the potential outcomes associated with each possible treatment trajectory with the *inverse probability of treatment weighted* (IPTW) estimator [175]. Specifically, a weight is assigned to each observation proportional to the inverse of the probability of treatment received given time-dependent confounders and previous treatments. The prediction models of these methods are typically based on linear or logistic regression. Therefore, in case the outcome or the treatment policy exhibits complex dependencies on the covariate history, it would output an inaccurate result. To overcome this problem, Lim et al. [99] proposed a deep learning method, *Recurrent Marginal Structural Networks* (R-MSN), to learn time-dependent effects by using an RNN architecture for treatment response estimation based on the MSM framework. Wodtke [175] suggests an alternative method for estimating marginal effects using *regression-with-residuals* (RWR) estimation of a constrained structural nested mean models. Unlike IPTW, this method does not require a model for the conditional probability of treatment at each time point making it a good candidate for continuous-valued treatment problems.

Another category of approaches to capture the treatment effect of the longitudinal data with time-varying treatment is the *Bayesian nonparametric* (BNP) methods. Xu et al. [179] used a flexible Bayesian nonparametric approach to estimate the disease trajectories as well as the univariate treatment response curves from sparse observational time series. Inspired by this work, Soleimani et al. [153] used the flexible Bayesian semiparametric approach and linear time-invariant dynamical systems to model the treatment effects in multivariate longitudinal data by capturing the dynamic behavior (i.e., time-varying treatments). Their model composed of three components, one that captures the treatment response and the other two components models the natural evolution of the signal independent of the treatment.

Most of the aforementioned methods assume that all the confounders are observed. In other words, they consider that variables affecting the treatment assignments and the potential outcomes are all known; otherwise, it will lead to a biased estimate of the outcome [26,106]. A natural way to overcome this problem is through *Factor Models* [36,55,143,180]. A recent deconfounder method introduced by Bica et al. [26] estimates the individual response to treatments in the presence of the multi-cause hidden confounders. To capture the distribution of the treatments over time, a factor model was built to create the latent variables. Moreover, to make sure that this factor model is able to estimate the distribution of the assigned causes,

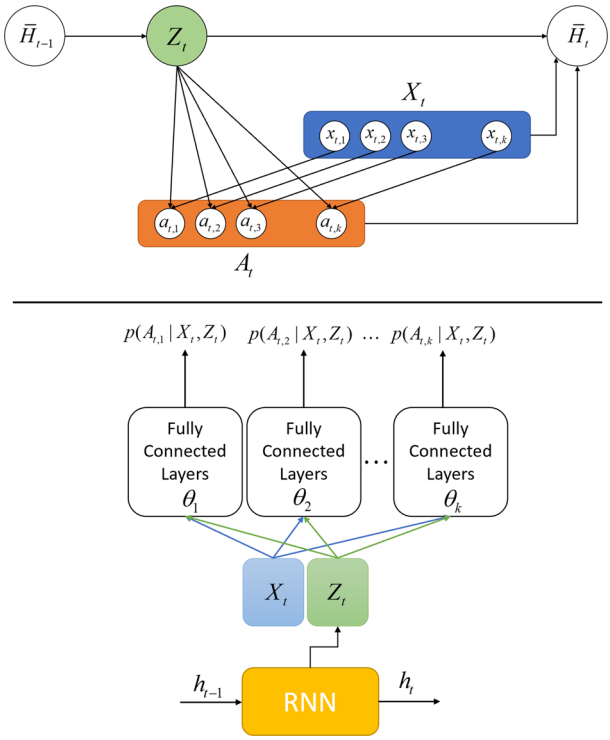


Fig. 3 (top) The factor model that creates the latent variables $z(t) = g(\bar{h}(t-1))$ where $\bar{h}(t-1)$ is the realization of history $\bar{H}(t-1)$. $X(t)$ is the covariates and $A(t)$ is the possible assignment of k treatment at time step t. (down) The implementation of the factor model using RNN [26]

a validation set of subjects were considered in order to compare the similarity of the two test statistics. These hidden confounders modeled by the latent variables were calculated by Recurrent Neural Network with multitask output and variational dropout (Fig. 3). Similarly, in the presence of hidden confounders, Liu et al. [105] proposed *Deep Sequential Weighting* (DSW) for estimating the ITE with time-varying confounders. The authors used deep recurrent weighting neural network to combine the current treatment assignments with historical information and learn a new representation of hidden confonders to predict the potential outcome.

4.1.3 Dynamic treatment regimes

With rapid-increasing interest in providing personalized treatment suggestions, dynamic treatment regimes are designed to provide treatment to individuals only when they need the treatment. A dynamic treatment regime is a function which takes in treatment and covariate history as arguments and outputs an action to be taken, providing a list of decision rules for how treatment should be allocated over time. Figure 4 shows a two-stage dynamic treatment regimes, where X and A denote the categorical covariates and the treatment, respectively. The observable data trajectory for a participant in a two-stage treatments is denoted by (X_1, A_1, X_2, A_2), where X_1 is the pre-treatment covariates and X_2 is the time-varying covariates which may depend on treatment received in the first interval. The randomized treatment actions are A_1 and A_2 and the primary outcome is $Y = f(X_1, A_1, X_2, A_2)$. For example, $X_2(A_1)$ denote a person's potential covariate status at the beginning of the second

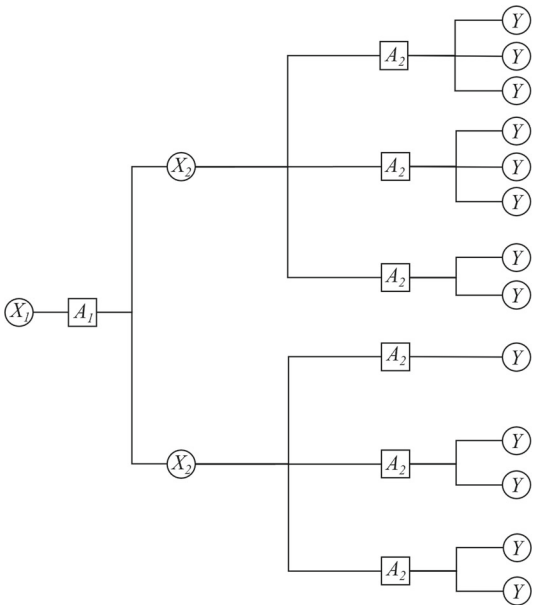


Fig. 4 Dynamic treatment regimes for two intervals. X and A are categorical data, hence, the tree-structured illustration [116]

interval if treatment A_1 is received by that person and Y denote the potential outcome if follows regime (A_1, A_2).

Regret function is widely used in dynamic treatment regimes to estimate the effect by the large scale treatments. To learn the parameters ψ in regret function, Moodie et al. [116] used g-estimation, which is proposed by Robins [132]. For the purpose of estimation, $S_j(a_j) = s_j(a_j, h_j)$ depends on variables which are considered as interaction with treatment to influence outcome, where h_j is unmeasured confounders. For example, if the function at the second interval is linear, $\gamma_2(h_2, a_2) = a_2(\psi_0 + \psi_1 x_1 + \psi_2 a_2 + \psi_3 x_2 a_1)$, then the analyst may choose $S_2(a_2) = \frac{\partial}{\partial \psi}\gamma_2(h_2, a_2) = a_2(1, x_2, a_1, x_2 a_1)^T$. Let

$$U(\psi, s) = \sum_{j_1}^{2} h_j(\psi) S_j(a_j) - E[S_j(a_j)|H_j], \tag{29}$$

with the probability of being the treated model. $E[U(\psi, s)] = 0$ is an unbiased estimating equation from which consistent estimates $\hat{\psi}$ of ψ may be found.

To avoid estimating the full multivariate distribution of the longitudinal data, Murphy et al. [118] design the regret function by modeling the multivariate distribution with two groups of parameters. The first group of parameters (parameters in the regret functions) will be estimated and used to estimate the optimal rules. The second group of parameters (most of which are infinite dimensional) are nuisance parameters. There are also some methods dealing with dynamic treatments by extending the conventional effect estimation methods. For example, Zhu et al. [184] extend Boruvka et al.'s [27] definition of lagged treatment effect to continuous treatments to define a weighted advantage function and proposed a novel SNM.

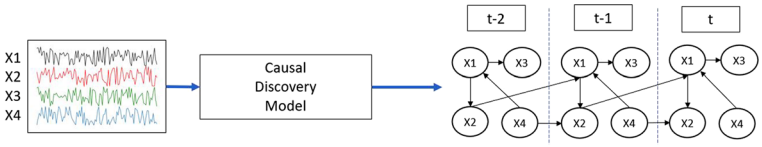


Fig. 5 A simple example showing how causal discovery models work by taking multivariate time series as input and output the causal relations between each series. This figure shows that $X2$ at time $t-2$ is a causal factor of $X1$ at time $t-1$ and $X4$ at timestep $t-2$ is a causal factor of $X2$ at timestep $t-1$

4.2 Causal discovery for time series

One of the fundamental tasks in any field of science is to identify the causal relations between different phenomena in a system [61,124,156]. Researchers are often interested in discovering what causes a phenomenon and how manipulating a phenomenon affects the others. Recently, there has been a proclivity toward creating algorithms for causal discovery on time series data [77,108,178]. Causal discovery on time series data is an important task and is often used in different fields of research. For example, determining the causal relations between the aggregated daily stock price and trading volumes or discovering how patient's records and the prescription of a specific drug over time are related to each other are types of questions which can be answered by performing causal discovery on time series data. Figure 5 shows an example of how causal discovery models work. In this section, we provide a comprehensive survey of the frameworks proposed for this task by categorizing them into three main types namely Granger causality and conditional independence-based approaches, structural equation-based models and deep learning-based frameworks and introducing the methods in each category.

4.2.1 Methods based on Granger causality and conditional independence

Granger causality [58] is a popular concept of causality which has been widely used to infer the causal relationships from time series data [11,45,70]. The idea behind Granger causality is that Y Granger causes X if it contains some unique information about X which is not available in X's past as well as all the information in the universe. In practice, this idea is materialized by investigating whether the prediction of the current value of time series X improves by incorporating Y's past into its own past. If so, it is reported that Y Granger causes X and has a causal influence on it. Vector autoregressive models (VAR) are common ways of modeling this problem:

$$X_t = \sum_{\tau=1}^{\tau_{max}} \phi(\tau) X_{t-\tau} + e_t,$$

where $X_t = (X_{1t}, ..., X_{nt})$ indicates time series X at time step t, $\phi(\tau)$ is the $N \times N$ coefficient matrix at lag τ , τ_{max} denotes the maximum time lag, and e represents an independent noise. Using this equation, we say X_i Granger causes X_j with lag τ if any of the coefficients in $\phi_{ji}(\tau)$ is nonzero. This relationship can be shown by $X^i_{t-\tau} \rightarrow X^j_t$ which demonstrates the causal link between X_i and X_j at lag τ.

Several VAR-based frameworks have been proposed over the years to perform causal discovery on time series. For instance, in [60], authors focus on linear vector autoregressive models along with stationary time series. They claim that adopting VAR model directly to

identify the causal relations neglects the possibility of shared structures in the lagged dependencies captured by the causal graph. To address this issue, they propose two new methods based on multi-task learning paradigms and techniques of structured regularizations for learning the G-causality in VARs with leading indicators. The difference between the two proposed methods is that they follow different structural assumptions for the G-causality graphs. The first method called SingleCluster VAR (SCVAR) assumes that the leading indicators within the system help predict all the series in the system . This method therefore aims at identifying such indicators; the second method named MultiCluster VAR (MCVAR) assumes that there are different indicators for different cluster of series and aims to learn the leading indicators as well as the unknown clusters.

Shojaie and Michailidis [151] propose a method for estimating the causal relationships of the time series for the case where number of components (p) is very large compared to the sample size (n). The authors claim that in such cases penalized methods provide higher accuracy. Based on this hypothesis, an extension of the lasso penalty named as truncating lasso penalty is proposed. This framework has two key features. First, it automatically determines the order of the Vector Autoregressive (VAR) models, and secondly, it performs model simplification by reducing the number of covariates in the model. The truncating lasso estimate of the graphical Granger model can be found by solving the following estimation problem for $i = 1, ..., p$:

$$\begin{aligned} &\underset{\theta^t \in \mathbb{R}^p}{\operatorname{argmin}} n^{-1} \left\| \mathcal{X}_i^T - \sum_{t=1}^{d} \mathcal{X}^{T-t} \theta^t \right\|_2^2 + \lambda \sum_{t=1}^{d} \Psi^t \sum_{j=1}^{p} \left| \theta_j^t \right| w_j^t \\ &\Psi^1 = 1, \quad \Psi^t = M^{I\{\|A^{(t-1)}\|_0 < p^2\beta/(T-t)\}}, t \geq 2, \end{aligned} \tag{30}$$

where $\mathcal{X}_i^t$ is the i-th column of the design matrix corresponding to t-th time point, M is a large constant, β is the allowed false negative rate set by the user and θ is the model parameters.

In [152], authors propose a framework to estimate the coefficients of VAR model when there are many observed variables and short time series. To tackle this scenario, the authors propose a new dimension reduction approach designed for time series. They modify the backward-in-time-selection (BTS) approach for Granger causality. While the original BTS method includes all lags up to the selected order p_k for each time series, X_k, the proposed modification includes only the lags of each X_k that are selected at each step of the algorithm.

In [6], authors claim that most of the existing works in inferring causal relations from time series data using Granger causality assume that the lag between a cause and an effect is at a fixed time point. To address this problem, they propose a novel method that uses Dynamic Time Warping (DTW) which is a distance measure between two time series along with Granger causality to identify the variable lagged-based causality in time series.

Haufe et al. [63] introduce a framework which aims to estimate causal interactions for multivariate time series. The authors argue that it is more practical to find all potential causal relations between all time series at once, rather than finding the causal relations for each time series pair. They propose a framework that accounts for the fact that there is no causal relation between z_i and z_j if all the AR coefficients for this certain pair of time series are jointly zero. They propose a sparsification technique through statistical testing.

Another way to recover the causal dependencies from times series data is by testing conditional independence relations between variables and their pasts. Conditional independence-based frameworks allow the causal graph identification under the assumption of time-order, Causal Sufficiency, Causal Markov Condition and Faithfulness. For example, transfer entropy [148] is an information theoretic approach for causal discovery which can

also be considered as the generalized Granger causality [20]. This method checks for conditional independence between $X^i_{t-\tau}$ and X^j_t given the past $(X_{t-1}, ..., X_{t-\tau})$. One major drawback of this method is that it highly suffers from curse of dimensionality. To alleviate this issue, Sun and Bollt [159] and Sun et al. [160] propose a framework based on optimal causation entropy principle. The framework identifies the lagged parents using an iterative procedure which consists of a forward and a backward phase.

Another conditional independence-based algorithm for causal discovery is PC algorithm [155]. Even though the original PC algorithm was proposed for general random variables and did not consider the time order of the them, the algorithm can be extended to the scenario where variables are collected chronologically. Entner and Hoyer propose to adopt the Fast Causal Inference (FCI) [154], originally designed for non-temporal data, to infer the causal relations from time series data in the presence of unobserved variables [48]. The advantage of the proposed method over Granger causality is that it also takes the latent variables in to account while identifying causal relations. Recently, a variation of PC algorithm, namely PCMCI algorithm [138], has also been proposed which consists of two steps. In the first step, the algorithm adopts PC algorithm to identify parents of X^j_t ($P(X^j_t)$) for all $X^j_t \in X_t$. It then applies a momentary conditional independence (MCI) test. In [140], the authors propose an extension to the PCMCI algorithm called PCMCI$^+$ that includes discovery of contemporaneous links along with lagged links.

Chu et al. [38] propose a causal discovery algorithm based on conditional independence which is specifically designed for nonlinear and stationary time series. The authors introduce two classes of conditional independence tests based on additive model regression to recover the structure of additive nonlinear time series consistently.

Jangyodsuk et al. [82] propose a novel approach for causal discovery in time series data based on Pearl's causality where the causal relationships are based on Directed Acyclic Graphs (DAGs) and the conditional dependencies between the variables. The authors state that the mutual information between an effect and a cause can be incrementally constructed based on the mutual information values between the effect Y and the previously found cases $X_{1:i-1}$. The final output of this method is a causal graph with each time series as a node and the edge weight for each edge is the leading time. Leading time is the difference in time steps between the occurrence of cause and the occurrence of the effect. The equation for mutual information chain rule is as shown:

$$I\left(X_{i:N}; Y\right) = \sum_{i=i}^{N} I\left(X_i; Y | X_{1:i-1}\right),$$

where Y is the effect and $X_i : N$ are its N causes.

4.2.2 Methods based on structural equation models

Structural Equation Models (SEM) have been long used to perform causal discovery from observational data [34,72,150,182]. One popular form of SEM used for identifying the causal structure is Linear Non-Gaussian Acyclic Model (LiNGAM) [150]. This model makes use of Independent Component Analysis (ICA) to identify causal relations in observational data. Unlike Gaussian Processes, LiNGAM is based on using the non-Gaussianity of the data. The key aspect of this method is that it is possible to identify more of the generating structure when the data are non-Gaussian. It can be mathematically represented as:

$$x_i = \sum_{k(j)<k(i)} b_{ij} x_j + e_i, \tag{31}$$

where e_i are continuous latent variables that are exogenous and b_{ij} are the connection strengths between variables x_i and x_j and k(i) denotes the causal ordering of x_i. The exogenous variables, e_i, follow Non-Gaussian distributions. In this section, we briefly introduce structural equation-based methods for causal discovery specifically designed for time series data.

In [78], Hyvärinen et al. propose a generalized version of LiNGAM which can be considered as a combination of autoregressive models and non-Gaussian instantaneous models as defined in equation below:

$$x(t) = \sum_{\tau=0}^{k} B_\tau x(t-\tau)_e(t), \tag{32}$$

where $e_i(t)$ are random processes which model disturbance. This model allows estimating instantaneous effects (i.e., $X_t^i \rightarrow X_t^j$) as well as lagged effects which is shown to lead to a better estimation of causal structure. The parameters of the model are estimated using a 3-step process by estimating an autoregressive model, computing the residuals and performing LiNGAM analysis on them. The paper also shows that the model is identifiable.

Schaechtle et al. [147] propose to integrate Linear Non-Gaussian Acyclic Model (LiNGAM) with tensor analytic techniques to identify causal relationships from the high dimensional data.

Rothenhäusler et al. [137] extend LiNGAM to learn linear causal cyclic models in the presence of latent variables. The authors explore the setting where the equilibrium data of a model are observed which is characterized by a set of linear relations Eq.(31).

$$CP(\mathbf{B}) = \max_{(m_1,\ldots,m_\eta,m_{\eta+1})_{1<\eta\leq p} \text{ cycle}} \prod_{1\leq k\leq \eta} \left|\mathbf{B}_{m_{k+1},m_k}\right|, \tag{33}$$

where B is the connectivity matrix. For their experiments, they assume that the data in an environment j are equilibrium observations of the model

$$\mathbf{x}_j = \mathbf{B}\mathbf{x}_j + \mathbf{c}_j + \mathbf{e}_j, \tag{34}$$

where c_j is the random intervention shift. Given these locations, the interventions shift the variables by a value determined by c_j. $\mathbf{6}_{\mathbf{c},j}$ is a diagonal matrix which is equivalent to demanding that interventions at different variables are uncorrelated. The final output of the model is the estimated connectivity matrix $\hat{B}$.

In [73], authors propose a time-varying causal model to represent the underlying causal process in nonstationary time series. The authors assume that the time series in consideration are generated from the LiNGAM process. They further allow each causal coefficient $b_{ij,t}$ and noise variance $\sigma_{i,t}^2$ to change over time which are modelled using the following autoregressive models:

$$\begin{aligned} b_{ij,t} &= \alpha_{ij,0} + \sum_{p=1}^{p_l} \alpha_{ij,p} b_{ij,t-p} + \epsilon_{ij,t} \\ h_{i,t} &= \beta_{i,0} + \sum_{q=1}^{q_l} \beta_{i,q} h_{i,t-q} + \eta_{i,t}, \end{aligned} \tag{35}$$

where $\epsilon_{ij,t} \sim \mathcal{N}\left(0, w_{ij}\right)$, $\eta_{i,t}\ \mathcal{N}\left(0, v_i\right)$, and $h_{i,t} = \log\left(\sigma_{i,t}^2\right)$ models the volatility of the observed time series. The authors further state that the time-varying linear causal model is actually a specific type of nonlinear state-space model with respect to hidden variables b_{ij}

and h_i. The causal graph is determined from the sampled particles. To determine whether there is a causal edge from x_j to x_i, the authors check both the mean and the variance of $\hat{b}_{ij,t}$. Specifically, if both $\overline{\hat{b}}_{ij} = \frac{1}{T}\sum_{t=1}^{T}\hat{b}_{ij,t} < \alpha$ and $\frac{1}{T}\sum_{t=1}^{T}\left(\hat{b}_{ij,t} - \overline{\hat{b}}_{ij}\right)^2 < \alpha$, then there is no causal edge between x_j and x_i.

Peters et al. [123] propose a class of structural equation models called Time Series Models with Independent Noise (TiMINo) for identifying the causal structure of a time series. The authors provide both theoretical analysis which provides general identifiability results and practical solution which introduces an algorithm based on testing the independence of the residuals for the case where there are no feedback loops between time series.

4.2.3 Deep learning-based methods

In this section, we briefly review the causal discovery frameworks which utilize the power of deep neural networks to perform causal discovery on time series data and overcome the shortcomings of traditional time series causal discovery frameworks.

Methods on Granger causality are built upon linearity of time series. However, in real-world cases, dependencies among time series are usually nonlinear and ignoring such interactions could lead to inconsistent estimation of Granger causality [164]. To incorporate nonlinear interactions into Granger causality detection, Tank et al. [162] propose a class of nonlinear architectures such as Multi Layer Perceptron (MLP) and Recurrent Neural Network (RNN) in which, each time series is modeled using an MLP or RNN. The input to the nonlinear framework is the past lags of all series, and the output is the future value of a single series. The authors also leverage a group lasso penalty to further shrink the weights of the inputs to zero. Dang et al. [44] propose a deep learning based framework which consists of multiple customized gated recurrent units (GRUs) designed to discover nonlinear and inter-time series dependencies. More specifically, the paper introduces a new dual purpose recurrent neural network which models the lagged dependencies in each time series and leverages those to discover inter-timeseries dependencies. In [176], the authors introduce a novel minimum predictive information regularization method to infer causal relations from time series, allowing deep learning models to discover nonlinear causal relations. This work makes two assumptions regarding the causality. First is the causal sufficiency assumption which states that each time series $x^{(i)}$ can only be caused by the time series from $x^{(1)}, x^{(2)}, ...x^{(N)}$. The second assumption is the "causality in mean" assumption which states that the causal relations influence the mean value of other variables. Their model tries to answer the question, how much can $X_{t-1}^{(j)}$ be corrupted without making the prediction of $x_t^{(i)}$ noticeably worse? To do so, they take the input and add independent noise with learnable amplitudes and measure the extent of corruption by mutual information between input and corrupted output.

The risk for the mentioned situation can be given by:

$$R_{\mathbf{X},x^{(i)}}[f_\theta, \boldsymbol{\eta}] = \mathbb{E}_{\mathbf{X}_{t-1},x_t^{(i)},\boldsymbol{\epsilon}}\left[\left(x_t^{(i)} - f_\theta\left(\tilde{\mathbf{X}}_{t-1}^{(\boldsymbol{\eta})}\right)\right)^2\right] + \lambda \cdot \sum_{j=1}^{N} I\left(\tilde{X}_{t-1}^{(j)(\eta_j)}; X_{t-1}^{(j)}\right), \tag{36}$$

where $\tilde{\mathbf{X}}_{t-1}^{(\boldsymbol{\eta})} := \mathbf{X}_{t-1} + \boldsymbol{\eta} \odot \boldsymbol{\epsilon}$ (or element-wise, $\tilde{X}_{t-1}^{(j)(\eta_j)} := X_{t-1}^{(j)} + \eta_j \cdot \epsilon_j$, $j = 1, 2, \ldots N$) is the noise-corrupted inputs with learnable noise amplitudes $\eta_j \in R^{KM}$, and $\epsilon_j \sim N(\mathbf{0}, \mathbf{I})$.$\lambda > 0$ is a positive hyperparameter for the mutual information $I(\cdot, \cdot)$.

At the minimization of $R_{\mathbf{X},x^{(i)}}[f_\theta, \boldsymbol{\eta}]$, W_{ji} is defined as $W_{ji} = I\left(\tilde{X}_{t-1}^{(j)\left(\eta_j^*\right)}; X_{t-1}^{(j)}\right)$, which is known as minimum predictive information and it measures the predictive strength of time series j for predicting time series i, conditioned on all other observed time series.

Most existing methods in causal discovery rely on predefined kernels or data distributions. To relax such assumptions, Xu et al. [178] propose a scalable causal discovery algorithm based on deep neural network. The proposed framework consists of four modules which account for temporal nonlinearity, learning the causal graph, identifying the intervariable nonlinearity and performing prediction for X_t in the framework of Granger causality. In order to make the framework more scalable, the authors also propose to approximate the causal graph via k-rank matrix decomposition.

In [120], the authors propose a hypernetwork framework for learning the intrinsic causality between spatial and temporal attributes to enhance the prediction performance of spatial temporal networks. They show that the spatial characteristics have a huge influence on the temporal characteristics and in order to capture this influence they use a hypernetwork.

Most of the aforementioned approaches are based on causal sufficiency assumption and therefore do not consider unobserved confounders. In such cases, observed confounders can be taken into account by controlling for them using for instance conditional Granger test [98]. However, in most real-world scenarios we cannot expect to have all possible confounders measured. Meng [112] addresses the problem of unobserved confounder in nonlinear Granger causality-based methods by approximating the distribution of unobserved confounders using Variational autoencoder. This distribution is sampled to get the estimated confounders which are used in the Granger test.

5 Performance evaluation

In this section, we present an overview of the benchmark datasets and metrics used in time series and causal time series literature.

The datasets and metrics that are used for evaluation are based on the type of models that we are evaluating . The fact that causal time series analysis targets a different problem than traditional time series modeling means that there are a lot more datasets available for time series which do not have causal metadata.

5.1 Datasets

In this section, we briefly introduce some of the datasets used in causal inference literature on time series. First, we begin by introducing datasets for the traditional time series forecasting. The functionalities of these datasets cannot be exploited for learning causality in time series as some datasets may lack ground truth information or may lack other features necessary for causal inference. To deal with these issues, we introduce datasets specifically catered for the causal discovery problem as well as the treatment effect estimation problems.

5.1.1 Time series datasets

Traditionally, time series analysis uses a variety of data. In this section, we discuss some of the most commonly used datasets for traditional time series problem.

- **UCR Time Series Classification Archive**: The UCR Time Series Classification archive consists of more than 120 datasets. These datasets represent a classification problem. Thus, there is a class label for each item in every dataset. The UCR time series dataset is used in various publications [17,49,177].
- **Baydogan's Archive**: Baydogan's Archive consists of over a dozen diverse multivariate time series datasets from different applications such as speech recognition, activity recognition and medicine. This dataset is used and introduced in [22].
- **NYC Taxi Dataset**: This is a univariate time-series dataset containing the New York City (NYC) taxi demand from 2014–07-01 to 2015-01-31 with an observation of the number of passengers recorded every half hour containing 10320 timestamps. This dataset is recommended by Braei et al. [29].
- **Real Yahoo Services Network Traffic Dataset**: This is a univariate time-series dataset containing the traffic to Yahoo services. The anomalies are labeled by humans. This dataset consists of 67 different time-series each containing about 1400 timestamps. This dataset is recommended by Braei et al. [29].
- **Synthetic Yahoo Services Network Traffic Dataset**: This dataset consists of 100 synthetic univariate time-series data containing anomalies. Each time-series contains about 1421 timestamps. The anomalies have been inserted randomly therefore representing point anomalies. This dataset is recommended by Braei et al. [29].

As mentioned earlier, these datasets are fruitful when it comes to traditional time series problems, like classification and forecasting. For inferring causality, one may need access to treatment and control groups for treatment effect estimation and may need access to variables such as confounders to interpret correct causal relations between different variables. Thus, the traditional datasets cannot be used for inferring causality as they lack these features. We now introduce datasets relevant for treatment effect estimation and causal discovery problems.

5.1.2 Treatment effect estimation datasets

In this section, we discuss some of the commonly used real-world datasets for the treatment effects estimation problem.

- **MIMIC II/III Data** [83]: This dataset consists of data about patients in ICU. Various attributes about the patients are stored in this database. Examples include blood pressure, oxygen saturation, given medicine, as well as temporal attributes. Thus, it is a gold-mine for causal research. This dataset has been used by Bica et al. [26] to estimate treatment effects. The dialysis subset of this data has been used by Soleimani et al. for Treatment-Response modeling using counterfactual reasoning [153].
- **Advertisement Data**: Brodersen et al. [30] use advertisement data by Google to determine the causal impact. An advertisement campaign was analyzed by the authors where the product ads were placed alongside search results for a period time of 6 weeks. The ad data were used as an intervention to measure the impact on sale volume as the effect.
- **Geo experiment data** [87]: These data consist of an ad campaign as a treatment for half of the non-overlapping geo data in the sample. It was used by [95] for experiments in their approach.

- **Economic data for Spanish regions**: In a case study about the economic cost of conflict, [2] use economic data for Spanish regions to analyze the effects of terrorism. The authors use the per capita GDP of Basque over time to do their causal analysis. The data do not have a ground truth value.
- **California's Tobacco Control Program**: Another work looks at the effects of California's tobacco control program in their work using Synthetic Control Methods [1]. Here, they use annual state-level panel data. These data contain the per-capita cigarette sales from 1970–2000 for multiple US states. This is the time between which Proposition 99 was passed.
- **Air Quality Data** [15]: This dataset is used to study the effects of gasoline content on air quality. This dataset includes Ozone levels, minimum/maximum/mean temperatures, precipitation and snow information. The data go over the years 1989 till 2006.
- **Monetary Policy Data**: This dataset comes from three different sources. Quarterly GDP for Switzerland and Euro are taken from Eurostat. The monthly business confidence index and monthly consumer price index for Switzerland are taken from OECD. Monthly balance sheet data, monthly call money rate and monthly average exchange rate are taken from the Swiss National Bank. Pfister et al. [125] use this dataset to satisfy their goal of finding monthly causal predictors for log-returns of the Euro-Swiss franc exchange rate. The authors use data from the year 1999 to 2017.

5.1.3 Causal discovery datasets

Datasets for causal discovery range from economical data [56] to health data [147]. We discuss some of the commonly used real-world datasets.

- **US Manufacturing Growth Data**: This dataset consists of microeconomic data of growth rates of US manufacturing firms in terms of employment, sales, research & development (R&D) expenditure, and operating income, for the years 1973-2004. It can be used to identify the causal variables that affect the growth rate of a firm. It has been used in [48].
- **Diabetes Dataset**: This dataset consists of Diabetes patient records that were obtained from two sources: an automatic electronic recording device and paper records. [147] use it to deduce the ground truth causal graph.
- **Temperature Ozone Data**: This dataset consists of two variables, 72 points in time, 16 different places. The two variables are ozone and radiation with the assumed ground truth that radiation has an causal effect on ozone. This dataset was used by [56,117,147].
- **OHDNOAA Dataset**: This is a dataset by the Office of Hydrologic Development at the National Oceanic and Atmospheric Administration which consists of 32 hydrology related variables over several square areas for the USA. The data are collected at constant intervals of 6 hours and ranged from the years 1979 to 2008. It is used by [82].
- **Neural activity Dataset**: This dataset consists of real-time whole-brain imaging to record the neural activity of Caenorhabditis elegans. The dataset consists of 302 neurons and is generally used to identify which neurons are responsible for movement.
- **Human Motion Capture**: This dataset is from CMU MoCap database contains data about joint angles, body position from two subjects. The dataset contains 54 joint angles over 2024 time points. Tank et al. [162] use this dataset in their Causal Discovery work.
- **Traffic Prediction Dataset**: This dataset contains four months' worth of sensor data from Los Angeles, California. 207 sensors are placed for collecting this data. The location of

Table 3 Properties of synthetic datasets used in the literature

Model	Description	Datasets
Confounding	Multiple effects from the same cause	Bica et al. [26]
Nonlinear	Absence of a linear relationship between cause and effect	Financial time-series [89], Particle simulation [108]
Dynamic time dependence	Dependence on the time-lagged component(s) varies over time	Li et al. [95]
Chaotic	Small changes in parameters create large changes over time	Papana et al. [121], Khanna et al. [88]

each sensor in the form of GPS coordinates are also included in the dataset. It is used by [120].

- **Stock Indices Data**: Stock Indices are a source of data used in Causal work. Rothenhausler et al. use NASDAQ, S&P 500, and DAX indices for a period between 2000–2012 for their Causal Discovery work [137]. They create 74 blocks of data. Each block represents 61 consecutive days.

5.1.4 Synthetic datasets

Even though there exist real-world datasets to evaluate the performance of causal discovery and the treatment effect estimation frameworks, many researchers use synthetic datasets for the purpose of illustrating particular technical difficulties inherent to some causal models, e.g., Markov equivalence (several causal graphs are consistent with the same data) or existence of confounding variables. In the following section, we discuss some synthetic dataset generation methods and some works that use them. A summary of the synthetic models is presented in Table 3.

- **Confounding/ Common-cause Models**: One of the concerns in the causal literature is the existence of confounders. Several approaches in the literature propose data generation processes to model their existence. For example, Huang et al. use simulated datasets with a common cause and common effect [74]. The datasets contain noise variables and causal variables over time. Let $e(t)$ be the value of a variable e at time t. Let c represent the variable which causes e. Then:

$$e(t) = \sum_{c \in X} \sum_{i=1}^{n} I(c, e),$$

where $n = |T(c) \cap [t - s, t - r]|$ and $I(c, e)$ is the impact of c on e.
- **NonLinear Models**: Since many proposed frameworks are designed for nonlinear systems, several approaches have been developed to simulate nonlinear systems for the evaluation. An example of nonlinear model is the simulation used by Papana et al. [121]. In their work, they simulate a tri-variate system with linear and nonlinear relationships. There are various other works which use nonlinear models of data [108].
- **Dynamic Models**: Some models try to simulate a situation where the dependence of the variables varies over time in a nonlinear and non-exponential manner. An example is the model used by Li et al. [95] where there is a sinusoidal dependence:

$$X_t = \cos(X_{t-1} + X_{t-4}) + \log(|X_{t-6} - X_{t-10}| + 1) + \epsilon_t$$

For treatment effects, one of the examples of a model used in literature is defined as follows [87]:

$$X_{i,t} = \beta_{0,t} + \beta_{1,t} X_{i,pre} + \beta_{2,t} Z_{i,t} + T_i \mu_{i,t} \cos(X_{i,pre}) + v_{i,t},$$

where Z is the time-varying covariate, $v_{i,t}$ is the noise, $X_{i,pre}$ is the pre-treatment covariate, T_i is the treatment indicator and $\mu_{i,t}$ represents the state.

- **Chaotic Models**: Chaoticity is the ability of a model to deviate over different values of its hyperparameters. Chaoticity is usually expressed in terms of a Lorenz model. Peters et al. [121] propose a simulation approach for nonlinear data with this property. Variations of the Lorenz model is used in other places in causal inference for time series. The Lorenze-96 model is used in the work by Khanna et al. [88] where the authors mention it as a popular model for climate science.

5.2 Evaluation metrics

In this section, we go over different metrics used in the literature for measuring how well a model performs. We begin by covering the metrics used for traditional time series problems like forecasting and classification. We then move on to metrics designed for causal discovery followed by the metrics for treatment effect estimation problem.

5.2.1 Time series metrics

There are multiple time series metrics used in the literature [18,49,103]. In the following, we discuss some of the most common ones.

- **Accuracy**: One of the common metrics for validating time series classification models is accuracy. The accuracy is simply the percentage of samples that are correctly classified [49,103].
- **Mean/Median Error**: There are several variations of errors used in time series literature, particularly for time series forecasting [75] such as (Root) Mean Squared Error, Mean/Median Absolute Error, Mean/ Median Absolute Percentage Error, Symmetric Mean/ Median Absolute Percentage Error, Median Relative Absolute Error, Geometric Mean Relative Absolute Error, and Mean Absolute Scaled Error. All of these are intended to measure the differences between the forecasted time series and the ground truth.
- **Longest Common SubSequence**: The LCS [18] measures the similarity of two time series with a matching threshold θ. A threshold θ is used to determine the state of match or mismatch between two time series. If the Euclidean distance of two data points is not greater than θ, then the two data points are considered to be a match, else they are said to mismatch.
- **Edit Distance with Real Penalty**: In this metric, the distance between the two series is measured by looking at the number of operations required to change one series into the other series. A distance matrix is created which stores the distances between the two time series across multiple time points inside the time window. Gaps between points are penalized according to a user-provided reference value [37].
- **Euclidean Distance**: One of the simplest evaluation metrics for time series is the Euclidean Distance (ED) [79]. Given two series, the square root of the sum of squared distances between each time step is the Euclidean distance. It helps define the similarity between two independent time series and is suitable for applications that do not present direct or necessary correlation among distinct features.

- **Dynamic Time Warping**: The dynamic time warping (DTW) [24] is a metric for calculating the distance between two time series sequences. Suppose two sequences of time series $A = a_1, a_2 \ldots, a_m$ and $B = b_1, b_2, \ldots b_n$ are given then, let DTW(i, j) denote the DTW distance of $A_{1..i}$ and $B_{1..j}$. The distance function dis(a_i, b_j) represents the distance of a_i and b_j. The Euclidean distance is usually used to measure the distance between a_i and b_j.

5.2.2 Causal time series evaluation metrics

In this section, we discuss different metrics for the evaluation of causality in time series. We first introduce metrics for the causal discovery problem followed by the metrics for the causal treatment effect estimation. A summary of the metrics can be found in Table 4.

Treatment effect estimation metrics When we talk about metrics for treatment effect estimation, we refer to those metrics that give feedback on how successful a model was in estimating a specific value. For causal effect estimation, the metrics, which are used, give us information about how well the estimation compares to the ground truth.

- **MSE**: MSE and its variations (Root Mean Squared Error [120], Normalized Root Mean Squared Error [153]) are used commonly in causal effect estimation literature [26,87]. Unlike MSE for causal discovery, here MSE is used to compare the inferred series with the ground truth series by taking the average of the squared differences at each time step.
- **F-Test**: F-test is used in treatment effect estimation to asses treatment effect heterogeneity by examining the marginal variances of the treatment and control outcomes. It is defined as the ratio of variance for the treated group over the variance for the control group.
- **T-Test**: The t-test is a metric used to compare two sequences. In their work on causal inference with rare time series data, Kleinberg et al. [89] use an unpaired t-test to determine the significance of a cause within a certain time-range.

Causal discovery metrics Here, we explain different metrics used in the literature for causal discovery. Causal discovery mainly focuses on finding the causal relationships [48,138], so the metrics involved in this type of approach usually provide measures of correctly identified relationships. Below, we list some of the commonly used metrics for this task:

- **Structural Hamming Distance**: Structural Hamming Distance (SHD) is a metric used to compare a discovered causal graph with the ground truth. More specifically, SHD describes the number of changes that need to be made to a graph to turn it into the graph its being compared with. This is calculated by counting the missing edges, extra edges and incorrect edge direction between two graphs. To asses the performance of causal discovery methods, SHD takes as input two partially directed acyclic graphs (PDAGs) and outputs the count of edges that do not coincide using the aforementioned process.
- **True/False Positive Rate**: The True Positive Rate for both adjacencies (discovered neighbors) and arrowheads (direction of discovered causal relations) are defined as the ratio of common edges found in the predicted and the ground truth adjacencies over the number of edges in the ground truth graph. It takes as input the predicted adjacency matrix and the ground truth adjacency matrix to calculate the ratio. Similarly, the False Positive Rate refers to the ratio of common edges found in the predicted and ground truth adjacencies over the absolute difference of number of edges between ground truth and predicted adjacencies. These metrics have been used in [77,138,139,147].

Table 4 Causal time series metrics and their definition

	Metric Name	Notions	Definition	Correctness	Comparative	Significance	Commonly Used in
1	TPR	Let $G = (V, E_M)$ be the causal graph inferred by the model, where $V_M = \{v_1, v_2, ..., v_n\}$ represent the covariates (including time-lagged) and E_M represent the edges between the covariates if there is a direct causal dependence Similarly, let $G' = (V, E_{GT})$ represent the covariates and the edges in the ground truth and $E = \{e_1, e_2, ..., e_m\}$ represent all possible edges between V	$TPR = \sum_i \frac{e_i}{\lvert E_{GT}\rvert},\ e_i \in E_M \cap E_{GT}$	✓			Causal Discovery
2	FPR		$FPR = \sum_i \frac{e_i}{\lvert E - E_{GT}\rvert},\ e_i \in E_M \backslash E_{GT}$	✓			Causal Discovery
3	MSE	Let T be the inferred transition matrix and A be the adjacency matrix for the ground truth	$MSE = \frac{1}{\lvert T\rvert} \sum (T - A)^{\circ 2}$	✓			Both
6	Precision	Let TP and FP denote the True Positive and False Positive, respectively	$Precision = \frac{TP}{TP+FP}$	✓			Causal Discovery
7	Recall	Let TP and FN denote the True Positive and False Negative, respectively	$Recall = \frac{TP}{TP+FN}$	✓			Causal Discovery

Table 4 continued

	Metric Name	Notions	Definition	Correctness	Comparative	Significance	Commonly Used in
8	F1-Score	Let P and R be the precision and recall, respectively	$F_1 = \frac{2PR}{P+R}$	✓			Causal Discovery
9	F-Test	Let $RSS_0 = \sum_{t=1}^{T} \hat{e}_t$ and $RSS_1 = \sum_{t=1}^{T} \hat{u}_t$, where T is the length of the time series, $\hat{e}_t$ and $\hat{u}_t$ is the time dependent error for the null hypothesis and the alternative hypothesis, respectively	$F = \frac{(RSS_0 - RSS_1)/p}{RSS_1/(T-2p-1)}$		✓	✓	Both
10	Unpaired T-Test	Let $\bar{x}_1$ and $\bar{x}_2$ be the means of two sequences. Let s_1 and s_2 represent the standard deviation of the sequences. Let n_1 and n_2 represent the cardinalities of the sequences	$utt = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)})}}$			✓	Treatment Effect Estimation

- **Area Under the Receiver Operator Curve**: This measure is one of the most popular metrics for Causal Discovery. The Receiver Operator Characteristic (ROC) is defined as a ratio of True Positive Rate (TPR) and False Positive Rate (FPR). The ROC curve is created by iterating over the cut off for classification and recording the TPR against the FPR. The area under the ROC curve (AUROC) is then used to assess the performance of the model. The higher the value of this metric, the better the model is. This metric has been widely used in different papers such as [63,88,108,176,178]. For Instance, in [88], the ROC curve illustrates the trade off between the TPR and the FPR achieved by the different methods toward the detection of pairwise Granger causal relationships between the n measured processes in their experiment.
- **Mean Squared Error**: This metric is used to evaluate works in causal discovery and for example, in Temporally Aggregated Time Series [56], the authors construct a transition matrix to represent the causal graph. Here, the MSE between the distance from the transition matrix to the ground truth is used to evaluate the causal relationships inferred by the proposed model.
- **F-Score, Precision and Recall**: Precision is defined as the ratio of correctly predicted positive observations to the total predicted positive observations. Recall is defined as the ratio of correctly predicted positive observations to the all observations in the actual class. F1 Score is defined as the weighted average of Precision and Recall. They are commonly used to evaluate the performance of a model. It takes as input the predicted adjacency matrix and the ground truth adjacency matrix to calculate the true positives, false negatives, true negatives and false negatives.
- **Area Under the Precision Recall Curve**: Area Under the Precision Recall Curve (AUPR) is another metric used in the literature [162,178]. This metric relies on precision and recall. Similar to the AUROC, this metric measures the area under the Precision-Recall curve.
- **F-Test**: One approach to finding whether causality exists is the F-test. To use this test, the user first defines a null hypothesis, for example in [121] the null hypothesis is, the coefficients of the lagged driving variables in the unrestricted VAR model are all zero . Then, we can construct a restricted and unrestricted equation and estimate the hyper-parameters using Ordinary Least Squares.

Definition 5 (F-Test) Let $X = \{x_t, x_{t-1}, x_{t-2}, ...\}$ be one series. Let $Y = \{y_t, y_{t-1}, y_{t-2}, ...\}$ be another series. The null hypothesis is given by the restricted equation:

$$x_t = c_1 + \sum_{i=1}^{p} \gamma_i x_{t-i} + e_t, \tag{37}$$

where p is the lag, c_1 is the intercept and e_t is the time-dependent error. The unrestricted equation represents the alternative hypothesis:

$$x_t = c_2 + \sum_{i=1}^{p} \alpha_i x_{t-i} + \sum_{i=1}^{p} \beta_i y_{t-i} + u_t, \tag{38}$$

where p is the lag, c_2 is the intercept and u_t is the time-dependent error. Let $RSS_0 = \sum_{t=1}^{T} \hat{e}_t$ and $RSS_1 = \sum_{t=1}^{T} \hat{u}_t$, where T is the length of the time series. The F-test statistic is defined as:

$$F = \frac{(RSS_0 - RSS_1)/p}{RSS_1/(T - 2p - 1)}. \tag{39}$$

6 Conclusion and future work

In this paper, we provide a comprehensive survey of causal inference tasks for time series data. We first categorize traditional ways of modeling this type of data into four categories, i.e., autoregressive models, dynamic Bayesian networks, Gaussian Processes and neural networks and discuss state-of-the-arts in each category. We then discuss two of the most important causal inference tasks on time series, i.e., causal treatment effect estimation and causal discovery. Each of these tasks is classified based on the type of approaches proposed to solve the problem and for each category, and state-of-the-arts are discussed. We also provide an extensive list of datasets and evaluation metrics used to assess the performance of frameworks proposed for different tasks. These metrics and datasets can be used as a guideline for future research in this field. Lastly, in the following, we discuss some future research opportunities in terms of estimating treatment effects, causal discovery and evaluating the performance of the models for each task.

6.1 Causal treatment effect estimation

Most of the methods in this task are base on the Stable Unit Treatment Values Assumption (SUTVA) which is that the potential outcome is not affected by exposure to the treatment of other units, and there is no hidden variations of treatment. However, in many fields of research such as social science, friends, families and acquaintances have influence on subjects' awareness of treatment and their desire to follow it [81]. Hence, there is a need to provide methods that account for social network and peer-influence both in participation decisions and in determining a subject's outcomes. Moreover, existing methods usually model treatments as discrete events. However, some treatments, for example dialysis or intravenous diuretics [153], are carried out continuously over a period of time. Therefore, designing methods for estimating the effect of continuous-time and continuous-valued treatments is a direction that needs to be explored.

subsectionCausal discovery Most exiting methods in causal discovery, especially methods based on deep neural networks, rely on the concept of Granger causality. Pearl's causality is another concept of causality which has been widely popular in i.i.d data scenarios [47,66,111]. One promising direction to pursue in future research on causal discovery for time series data is to utilize the power of deep neural networks to learn the causal structure from time series data. Moreover, most exiting algorithms in time series field only leverage observational data. Interventional data have proven to be extremely useful in learning more accurate causal structures from i.i.d data [31,41,157]. We therefore suggest leveraging both observational and interventional time series data to learn a better and more accurate causal structure .

6.2 Performance evaluation and benchmark datasets

In terms of data, both treatment effect estimation and causal discovery tasks need to include more robust datasets for evaluation purposes. For example, in causal discovery, we need datasets to evaluate multi-modal causal discovery algorithms when one of the modalities is time. Multimodal data have gained a lot of attention in various fields like healthcare [10] and financial [93], thus promoting the need for such datasets for evaluation. Multimodal data represent data of different types like images, text, etc. For example, satellite images of a scene taken over different time represents multimodal data. When it comes to treatment effect estimation, there is a need for datasets that are suitable for calculating the counter-

factual outcomes along with the factual outcomes as they help in estimating the individual treatment effects which would help personalize the effect of a treatment on each individual rather than considering an average population. Most of the datasets mentioned in the earlier section are generated from observational studies and its impossible to obtain both factual and counterfactual outcomes in such a setting. There have been works like [80,107], that use constructed data which is a mix of observational and randomized control trials (RCT) to overcome this difficulty. We need more such data for better evaluation of treatment effect estimation.

References

1. Abadie A, Diamond A, Hainmueller J (2010) Synthetic control methods for comparative case studies: estimating the effect of Californias tobacco control program. J Am Stat Assoc 105(490):493–505
2. Abadie A, Gardeazabal J (2003) The economic costs of conflict: a case study of the Basque Country. Am Econ Rev 93(1):113–132
3. Abanda A, Mori U, Lozano JA (2019) A review on distance based time series classification. Data Min Knowl Dis 33(2):378–412
4. Abrevaya J, Hsu YC, Lieli RP (2015) Estimating conditional average treatment effects. J Bus Econ Stat 33(4):485–505
5. Amjad M, Shah D, Shen D (2018) Robust synthetic control. J Mach Learn Res 19(1):802–852
6. Amornbunchornvej C, Zheleva E, Berger-Wolf TY (2019) Variable-lag Granger Causality for Time Series Analysis. In: 2019 IEEE international conference on data science and advanced analytics (DSAA). IEEE, pp 21–30
7. Andrieu C, Doucet A, Holenstein R (2010) Particle Markov chain Monte Carlo methods
8. Angrist JD, Pischke JS (2008) Mostly harmless econometrics: an empiricists companion. Princeton University Press, Princeton
9. Angrist JD, Pischke JS (2014) Masteringmetrics: the path from cause to effect. Princeton University Press, Princeton
10. Anwar AR et al (2014) Multi-modal causality analysis of eyes-open and eyes-closed data from simultaneously recorded EEG and MEG. In: 2014 36th annual international conference of the IEEE engineering in medicine and biology society. IEEE, pp 2825–2828
11. Arnold A, Liu Y, Abe N (2007) Temporal causal modeling with graphical granger methods. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. pp 66–75
12. Atanasov VA, Black BS (2016) Shock-based causal inference in corporate finance and accounting research. Crit Financ Rev 5:207–304
13. Athey S, Imbens GW (2006) Identification and inference in nonlinear difference-in-differences models. Econometrica 74(2):431–497
14. Athey S, Imbens GW (2017) The state of applied econometrics: causality and policy evaluation. J Econ Perspec 31(2):3–32
15. Auffhammer M, Kellogg R (2011) Clearing the air? The effects of gasoline content regulation on air quality. Am Econ Rev 101(6):2687–2722
16. Aytuğ H et al (2017) Twenty years of the EU-Turkey customs union: a synthetic control method analysis. JCMS J Common Market Stud 55(3):419–431
17. Bagnall A, Lines J, Hills J, Bostrom A (2015) Time-series classification with COTE: the collective of transformation-based ensembles. IEEE Trans Knowl Data Eng 27(09):1. https://doi.org/10.1109/TKDE.2015.2416723
18. Bagnall A et al (n.d) The great time series classification bake off: An experimental evaluation of recently proposed algorithms. Extended version. arXiv 2016. In: arXiv preprint arXiv:1602.01711
19. Balzer LB, Petersen ML, van der Laan MJ, Search Collaboration (2016) Targeted estimation and inference for the sample average treatment effect in trials with and without pair-matching. Stat Med 35(21):3717–3732
20. Barnett L, Barrett AB, Seth AK (2009) Granger causality and transfer entropy are equivalent for Gaussian variables. Phys Rev Lett 103(23):238701
21. Baum LE et al (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. Ann Math Stat 41(1):164–171. https://doi.org/10.1214/aoms/1177697196

22. Baydogan MG, Runger G (2015) Learning a symbolic representation for multivariate time series classification. Data Min Knowl Dis 29(2):400–422
23. Bernal JL, Cummins S, Gasparrini A (2017) Interrupted time series regression for the evaluation of public health interventions: a tutorial. Int J Epidemiol 46(1):348–355
24. Berndt DJ, Clifford J (1994) Using dynamic time warping to find patterns in time series. In: Proceedings of the 3rd international conference on knowledge discovery and data mining. AAAIWS'94. AAAI Press, Seattle, WA pp 359–370
25. Bertrand M, Duflo E, Mullainathan S (2004) How much should we trust differences-in-differences estimates? Q J Econ 119(1):249–275
26. Bica I, Alaa AM, van der Schaar M (2019) Time series deconfounder: estimating treatment effects over time in the presence of hidden confounders. In: arXiv preprint arXiv:1902.00450
27. Boruvka A, Almirall D, Witkiewitz K, Murphy SA (2018) Assessing time-varying causal effect moderation in mobile health. J Am Stat Assoc 113(523):1112–1121
28. Box GEP, Jenkins GM (1968) Some recent advances in forecasting and control. J R Stat Soc Series C 17(2):91–109. https://doi.org/10.2307/2985674
29. Braei M, Wagner S (2020) Anomaly Detection in Univariate Time-series: A Survey on the State-of-the-Art
30. Brodersen KH, Gallusser F, Koehler J, Remy N, Scott SL (2015) Inferring causal impact using Bayesian structural time-series models. Ann Appl Stat 9(1):247–274
31. Brouillard P et al (2020) Differentiable causal discovery from interventional data. In: arXiv preprint arXiv:2007.01754
32. Brown RG (1956) Exponential smoothing for predicting demand. Little https://books.google.com/books?id=Eo_rMgEACAAJ
33. Bruhn CAW et al (2017) Estimating the population-level impact of vaccines using synthetic controls. Proc Natl Acad Sci 114(7):1524–1529
34. Cai R et al (2018) Causal discovery from discrete data using hidden compact representation In: Advances in neural information processing systems, p 2666
35. Cavallo E et al (2013) Catastrophic natural disasters and economic growth. Rev Econ Stat 95(5):1549–1561
36. Chan MK, Kwok S et al (2016) Policy evaluation with interactive fixed effects. In: Preprint. Available at https://ideas.repec.org/p/syd/wpaper/2016-11.html
37. Chen, L, Ng R (2004) On the marriage of lp-norms and edit distance. In: Proceedings of the thirtieth international conference on very large data bases, 30:792–803
38. Chu T, Glymour C, Ridgeway G (2008) Search for additive nonlinear time series causal models. J Mach Learn Res 9(5)
39. Cole MA, Elliott RJR, Liu B (2020) The impact of the Wuhan Covid-19 lockdown on air pollution and health: a machine learning and augmented synthetic control approach. Environ Res Econ 1–28
40. Cooley J, Navarro S, Takahashi Y (2010) Identification and estimation of time-varying treatment effects: How the timing of grade retention affects outcomes. In: manuscrit, University of Wisconsin-Madison
41. Cooper GF, Yoo C (2013) Causal discovery from a mixture of experimental and observational data. In: arXiv preprint arXiv:1301.6686
42. Cunningham J, Ghahramani Z, Rasmussen C (2012) Gaussian processes for time-marked time-series data In: Artificial intelligence and statistics, pp 255–263
43. Damianou A, Lawrence N (2013) Deep gaussian processes. In: Artificial intelligence and statistics, pp 207–215
44. Dang XH, Shah SY, Zerfos P (2018) seq2graph: discovering dynamic dependencies from multivariate time series with multi-level attention. In: arXiv preprint arXiv:1812.04448
45. Ding M, Chen Y, Bressler SL (2006) 17 Granger causality: basic theory and application to neuroscience. In: Handbook of time series analysis: recent theoretical developments and applications 437
46. Eichler M, Didelez V (2012) Causal reasoning in graphical time series models. In: arXiv preprint arXiv:1206.5246
47. Ellis B, Wong WH (2008) Learning causal Bayesian network structures from experimental data. J Am Stat Assoc 103(482):778–789
48. Entner D, Hoyer PO (2010) On causal discovery from time series data using FCI. In: Probabilistic graphical models, pp 121–128
49. Fawaz HI et al (2019) Deep learning for time series classification: a review. Data Min Knowl Dis 33(4):917–963
50. Fu T (2011) A review on time series data mining. Eng Appl Artif Intel 24(1):164–181. https://doi.org/10.1016/j.engappai.2010.09.007
51. Gamboa JCB (2017) Deep learning for time-series analysis. In: arXiv preprint arXiv:1701.01887

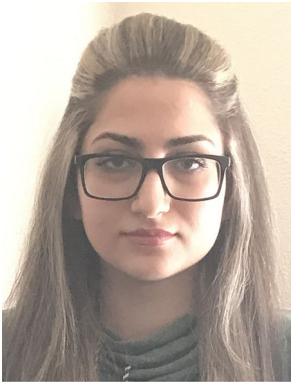
52. Ghahramani Z (1998) Learning Dynamic Bayesian Networks. In: In Adaptive processing of sequences and data structures, Lecture Notes in Artificial Intelligence, pp 168–197
53. Ghahramani Z, Hinton GE (1996) Switching State-Space Models. Tech. rep. Kings College Road, Toronto M5S 3H5
54. Ghahramani Z, Jordan MI (1996) Factorial Hidden Markov Models. In: Machine Learning, MIT Press
55. Gobillon L, Magnac T (2016) Regional policy evaluation: interactive fixed effects and synthetic controls. Rev Econ Stat 98(3):535–551
56. Gong M et al (2017) Causal discovery from temporally aggregated time series. In: Uncertainty in artificial intelligence: proceedings of the... conference. In: conference on uncertainty in artificial intelligence Vol. 2017. NIH Public Access
57. González R, Hosoda EB (2016) Environmental impact of aircraft emissions and aviation fuel tax in Japan. J Air Transp Manag 57:234–240
58. Granger CWJ (1969) Investigating causal relations by econometric models and cross-spectral methods. Econom J Econom Soc 37:424–438
59. Graves A (2013) Generating sequences with recurrent neural networks. In: CoRR arXiv:1308.0850
60. Gregorova M, Kalousis A, Marchand-Maillet S (2015) Learning Leading Indicators for Time Series Predictions. In: arXiv preprint arXiv:1507.01978
61. Guo R et al (2018) A survey of learning causality with data: problems and methods. In: arXiv preprint arXiv:1809.09337
62. HajiGhassemi N, Deisenroth M (2014) Analytic long-term forecasting with periodic Gaussian processes. In: Artificial Intelligence and Statistics, pp 303–311
63. Haufe S et al (2010) Sparse causal discovery in multivariate time series. In: causality: objectives and assessment, pp 97–106
64. Hausman C, Rapson DS (2018) Regression discontinuity in time: considerations for empirical applications. Annu Rev Res Econ 10:533–552
65. Marton H, Hernéndez-Lobato JM, Murillo-Fuentes JJ (2018) Inference in deep gaussian processes using stochastic gradient hamiltonian monte carlo. In: Advances in neural information processing systems, pp 7506–7516
66. Heckerman D (2013) A Bayesian approach to learning causal networks. In: arXiv preprint arXiv:1302.4958
67. Hedeker D, Gibbons RD (2006) Longitudinal data analysis, vol 451. Wiley, Hoboken
68. Hernán MA, Robins JM (2010) Causal inference
69. Hernán MA, Brumback B, Robins JM (2000) Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. In: Epidemiology, pp 561–570
70. Hiemstra C, Jones JD (1994) Testing for linear and nonlinear Granger causality in the stock price-volume relation. J Financ 49(5):1639–1664
71. Hinton GE (2002) Training products of experts by minimizing contrastive divergence. Neural Comput 14(8):1771–1800
72. Hoyer P et al (2008) Nonlinear causal discovery with additive noise models. Adv Neural Inf Process Syst 21:689–696
73. Huang B et al (2019) Causal discovery and forecasting in nonstationary environments with state-space models. Proc Mach Learn Res 97:2901
74. Huang Y, Kleinberg S (2015) Fast and accurate causal inference from time series data. In: The twenty-eighth international flairs conference
75. Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. Int J Forecast 22(4):679–688
76. Rob H et al (2002) A state space framework for automatic forecasting using exponential smoothing methods. Int J Forecast 18:439–454. https://doi.org/10.1016/S0169-2070(01)00110-8
77. Hyttinen A, Plis S, Järvisalo M, Eberhardt F, Danks D (2016) Causal discovery from subsampled time series data by constraint optimization. In: Conference on probabilistic graphical models. PMLR, pp 216–227
78. Hyvärinen A, Shimizu S, Hoyer PO (2008) Causal modelling combining instantaneous and lagged effects: an identifiable model based on non-Gaussianity. In: Proceedings of the 25th international conference on Machine learning, pp 424–431
79. Iglesias F, Kastner W (2013) Analysis of similarity measures in times series clustering for the discovery of building energy patterns. Energies 6(2):579–597
80. Jaber A et al (2020) Causal discovery from soft interventions with unknown targets: characterization and learning. In: Advances in neural information processing systems 33
81. Jackson MO, Lin Z, Yu NN (2020) Adjusting for peer-influence in propensity scoring when estimating treatment effects. In: Available at SSRN 3522256

82. Jangyodsuk P, Seo DJ, Gao J (2014) Causal graph discovery for hydrological time series knowledge discovery
83. Johnson AEW et al (2016) MIMIC-III, a freely accessible critical care database. Sci Data 3(1):1–9
84. Julier SJ, Uhlmann JK (1997) A new extension of the kalman filter to nonlinear systems. In: pp 182–193
85. Kalman RE et al (1960) A new approach to linear filtering and prediction problems. J Basic Eng 82(1):35–45
86. Karl M et al (2017) Deep variational bayes filters: unsupervised learning of state space models from raw data. arXiv:1605.06432 [stat.ML]
87. Kerman J, Wang P, Vaver J (2017) Estimating ad effectiveness using geo experiments in a time-based regression framework
88. Khanna S, Tan VYF (2019) Economy statistical recurrent units for inferring nonlinear granger causality. In: arXiv preprint arXiv:1911.09879
89. Kleinberg S (2013) Causal inference with rare events in large-scale time-series data. In: twenty-third international joint conference on artificial intelligence
90. Kontopantelis E et al (2015) Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. BMJ 350:h2750
91. Kreif N et al (2016) Examination of the synthetic control method for evaluating health policies with multiple treated units. Health Econ 25(12):1514–1528
92. Lechner M et al (2011) The estimation of causal effects by difference-in-difference methods. Now
93. Lee SI, Yoo SJ (2019) Multimodal deep learning for finance: integrating and forecasting international stock markets. J Supercomput 1–19
94. Li L, Prakash BA (2011) Time series clustering: complex is simpler!. In: ICML
95. Li S (2018) Estimating causal effects from time series. PhD thesis. ETH Zurich
96. Li S, Bühlmann P (2018) Estimating heterogeneous treatment effects in nonstationary time series with state-space models. In: arXiv preprint arXiv:1812.04063
97. Li SCX, Marlin B (2016) A scalable end-to-end gaussian process adapter for irregularly sampled time series classification. In: Advances in neural information processing systems, pp 1804–1812
98. Liao W et al (2010) Evaluating the effective connectivity of resting state networks using conditional Granger causality. Biol Cybern 102(1):57–69
99. Lim B (2018) Forecasting treatment responses over time using recurrent marginal structural networks. In: advances in neural information processing systems, pp 7483–7493
100. Lim B, Zohren S (2021) Time-series forecasting with deep learning: a survey. Philos Trans R Soc A 379(2194):20200209
101. Linden A, Adams JL (2011) Applying a propensity score-based weighting model to interrupted time series data: improving causal inference in programme evaluation. J Eval Clin Prac 17(6):1231–1238
102. Lines J, Bagnall A (2014) Ensembles of elastic distance measures for time series classification. In: Proceedings of the 2014 SIAM international conference on data mining. SIAM, pp 524–532
103. Lines J, Taylor S, Bagnall A (2016) Hive-cote: the hierarchical vote collective of transformation-based ensembles for time series classification. In: 2016 IEEE 16th international conference on data mining (ICDM). IEEE, pp 1041–1046
104. Liu H et al (2020) When Gaussian process meets big data: a review of scalable GPs. IEEE Trans Neural Netw Learn Syst 31(11):4405–4423
105. Liu R, Yin C, Zhang P (2020) Estimating individual treatment effects with time-varying confounders. In: arXiv preprint arXiv:2008.13620
106. Lok JJ et al (2008) Statistical modeling of causal effects in continuous time. Ann Stat 36(3):1464–1507
107. Louizos C et al (2017) Causal effect inference with deep latent-variable models. In: arXiv preprint arXiv:1705.08821
108. Löwe S et al (2020) Amortized causal discovery: learning to infer causal graphs from time-series data. In: arXiv preprint arXiv:2006.10833
109. Qianli M et al (2019) Learning representations for time series clustering. In: Wallach H et al (eds) Advances in neural information processing systems, vol 32. Curran Associates Inc, New York, pp 3781–3791
110. Maddix DC, Wang Y, Smola A (2018) Deep factors with gaussian processes for forecasting. In: arXiv preprint arXiv:1812.00098
111. Meganck S, Leray P, Manderick B (2006) Learning causal bayesian networks from observations and experiments: a decision theoretic approach. In: international conference on modeling decisions for artificial intelligence. Springer, pp 58–69
112. Meng Y (2019) Estimating granger causality with unobserved confounders via deep latent-variable recurrent neural network. In: arXiv preprint arXiv:1909.03704

113. Mittelman R (2015) Time-series modeling with undecimated fully convolutional neural networks. arXiv preprint arXiv:1508.00317
114. Mitze T, Kosfeld R, Rode J, Wälde K (2020) Face masks considerably reduce COVID-19 cases in Germany. Proc Nat Acad Sci 117(51):32293–32301
115. Mogren O (2016) C-RNN-GAN: Continuous recurrent neural networks with adversarial training. In: CoRR arXiv:1611.09904
116. Moodie EEM, Richardson TS, Stephens DA (2007) Demystifying optimal dynamic treatment regimes. Biometrics 63(2):447–455
117. Mooij JM et al (2016) Distinguishing cause from effect using observational data: methods and benchmarks. J Mach Learn Res 17(1):1103–1204
118. Murphy SA (2003) Optimal dynamic treatment regimes. J R Stat Soc Series B (Stat Methodol) 65(2):331–355
119. van den Oord A et al (2016) WaveNet: a generative model for raw audio. In arXiv:1609.03499
120. Pan Z et al (2018) Hyperst-net: Hypernetworks for spatio-temporal forecasting. In: arXiv preprint arXiv:1809.10889
121. Papana A et al (2013) Simulation study of direct causality measures in multivariate time series. Entropy 15(7):2635–2661
122. Penfold RB, Zhang F (2013) Use of interrupted time series analysis in evaluating health care quality improvements. Acad Pediatr 13(6):S38–S44
123. Peters J, Janzing D, Schölkopf B (2013) Causal inference on time series using restricted structural equation models. In: advances in neural information processing Systems, pp 154–162
124. Peters J, Janzing D, Schölkopf B (2017) Elements of causal inference. The MIT Press, Cambridge
125. Pfister N, Bühlmann P, Peters J (2019) Invariant causal prediction for sequential data. J Am Stat Assoc 114(527):1264–1276
126. Quiñonero-Candela J, Rasmussen CE (2005) A unifying view of sparse approximate Gaussian process regression. J Mach Learn Res 6(Dec):1939–1959
127. Rasmussen CE (2003) Gaussian processes in machine learning. In: summer school on machine learning. Springer, pp 63–71
128. Roberts PSS (2002) Bayesian time series classification. Adv Neural Inf Process Syst 14:937
129. Roberts S et al (2013) Gaussian processes for time-series modelling. Philos Trans R Soc A Math Phys Eng Sci 371(1984):20110550
130. Robins J (1992) Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. Biometrika 79(2):321–334
131. Robins JM (1997) Causal inference from complex longitudinal data. In: Latent variable modeling and applications to causality. Springer, pp 69–117
132. Robins JM (2004) Optimal structural nested models for optimal sequential decisions. In: Proceedings of the second seattle Symposium in Biostatistics. Springer, pp 189–326
133. Robins JM, Greenland S, Hu FC (1999) Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. J Am Stat Assoc 94(447):687–700
134. Robins JM, Hernan MA, Brumback B (2000) Marginal structural models and causal inference in epidemiology
135. Robins J, Hernan M (2008) Estimation of the causal effects of time-varying exposure. In: pp 553–599 https://doi.org/10.1201/9781420011579.ch23
136. Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55
137. Rothenhäusler D et al (2015) BACKSHIFT: Learning causal cyclic graphs from unknown shift interventions. In: advances in neural information processing systems, pp 1513–1521
138. Runge J, Sejdinovic D, Flaxman S, (n.d) Detecting causal associations in large nonlinear time series datasets. arXiv 2017. In: arXiv preprint arXiv:1702.07007
139. Runge J (2018) Causal network reconstruction from time series: from theoretical assumptions to practical estimation. Chaos Interdiscip J Nonlinear Sci 28(7):075310
140. Runge J (2020) Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In: arXiv preprint arXiv:2003.03685
141. Runge J et al (2019) Inferring causation from time series in Earth system sciences. Nat Commun 10(1):1–13
142. Saatçi Y (2012) Scalable inference for structured Gaussian process models. PhD thesis. Citeseer
143. Samartsidis P, Seaman SR, Montagna S, Charlett A, Hickman M, Angelis DD (2020) A bayesian multivariate factor analysis model for evaluating an intervention by using observational time series data on multiple outcomes. J Royal Stat Soc Series A (Statistics in Society) 183(4):1437–1459

144. Samartsidis P, Seaman SR, Presanis AM et al (2019) Assessing the causal effect of binary interventions from observational panel data with few treated units. Stat Sci 34(3):486–503
145. Saul LK, Jordan MI (1998) Mixed memory Markov models: decomposing complex stochastic processes as mixtures of simpler ones
146. Saunders J et al (2015) A synthetic control approach to evaluating place-based crime interventions. J Quant Criminol 31(3):413–434
147. Schaechtle U, Stathis K, Bromuri S (2013) Multi-dimensional causal discovery. In: twenty-third international joint conference on artificial intelligence
148. Schreiber T (2000) Measuring information transfer. Phys Rev Lett 85(2):461
149. Shannon M, Byrne W (2009) A formulation of the autoregressive HMM for speech synthesis
150. Shimizu S et al (2006) A linear non-Gaussian acyclic model for causal discovery. J Mach Learn Res 7(10):2003–2030
151. Shojaie A, Michailidis G (2010) Discovering graphical Granger causality using the truncating lasso penalty. Bioinformatics 26(18):i517–i523
152. Siggiridou E, Kugiumtzis D (2015) Granger causality in multivariate time series using a time-ordered restricted vector autoregressive model. IEEE Trans Signal Process 64(7):1759–1773
153. Soleimani H, Subbaswamy A, Saria S (2017) Treatment-response models for counterfactual reasoning with continuous-time, continuous-valued interventions. In: arXiv preprint arXiv:1704.02038
154. Spirtes P, Glymour C, Scheines R (2000) Causation, prediction, and search, 2nd edn. MIT Press, Cambridge MA
155. Spirtes P, Glymour C (1991) An algorithm for fast recovery of sparse causal graphs. Soc Sci Comput Rev 9(1):62–72
156. Spirtes P, Zhang K (2016) Causal discovery and inference: concepts and recent methodological advances. In: Applied informatics. vol. 3. 1. Springer, p. 3
157. Steyvers M et al (2003) Inferring causal networks from observations and interventions. Cognit Sci 27(3):453–489
158. Stips A et al (2016) On the causal structure between CO 2 and global temperature. Sci Rep 6(1):1–9
159. Sun J, Bollt EM (2014) Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings. Phys D Nonlinear Phenom 267:49–57
160. Sun J, Taylor D, Bollt EM (2015) Causal network inference by optimal causation entropy. SIAM J Appl Dyn Syst 14(1):73–106
161. Sutskever I, Hinton GE (2007) Learning Multilevel Distributed Representations for High-Dimensional Sequences. In: Meila M, Shen X (Eds.), AISTATS Vol 2. JMLR Proceedings. JMLR.org, pp 548–555. http://dblp.uni-trier.de/db/journals/jmlr/jmlrp2.html#SutskeverH07
162. Tank A et al (2018) Neural granger causality for nonlinear time series. In: arXiv preprint arXiv:1802.05842
163. Taylor GW (2009) Composable, distributed-state models for high-dimensional time series. University of Toronto, Toronto
164. Teräsvirta T, Tjøstheim D, Granger C et al (2010) Modelling nonlinear economic time series. Oxford University Press, Oxford
165. Tobar F, Bui TD, Turner RE (2015) Learning stationary time series using Gaussian processes with nonparametric kernels. In: Advances in neural information processing systems, pp 3501–3509
166. de Vocht F (2016) Inferring the 1985–2014 impact of mobile phone use on selected brain cancer subtypes using Bayesian structural time series and synthetic controls. Environ Int 97:100–107
167. de Vocht F et al (2017) The intervention effect of local alcohol licensing policies on hospital admission and crime: a natural experiment using a novel Bayesian syntheticstime-series method. J Epidemiol Commun Health 71(9):912–918
168. Wang JM, Fleet DJ, Hertzmann A (2006) Gaussian process dynamical models. In: In NIPS. MIT Press, pp 1441–1448
169. Wang Y et al (2019) Deep factors for forecasting. In: International conference on machine learning. PMLR, pp 6607–6617
170. Wilson AG et al (2016) Deep kernel learning. In: Artificial intelligence and statistics, pp 370–378
171. Wilson AG et al (2016) Stochastic variational deep kernel learning. Adv Neural Inf Process Syst 29:2586–2594
172. Wilson A, Adams R (2013) Gaussian process kernels for pattern discovery and extrapolation. In: International conference on machine learning, pp 1067–1075
173. Wilson A, Nickisch H (2015) Kernel interpolation for scalable structured Gaussian processes (KISSGP). In: international conference on machine learning, pp 1775–1784
174. Wing C, Simon K, Bello-Gomez RA (2018) Designing difference in difference studies: best practices for public health policy research. In: Annual review of public health 39

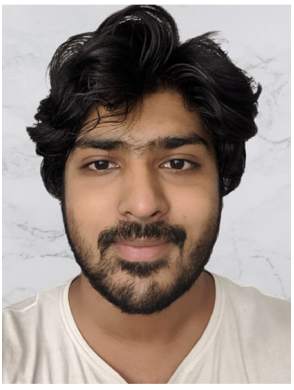
175. Wodtke GT (2020) Regression-based adjustment for time-varying confounders. Sociol Methods Res 49(4):906–946
176. Wu T, Breuel T, Skuhersky M, Kautz J. Nonlinear causal discovery with minimum predictive information regularization
177. Xing Z, Pei J, Keogh E (2010) A brief survey on sequence classification. ACM Sigkdd Explor Newsl 12(1):40–48
178. Xu C, Huang H, Yoo S (2019) Scalable causal graph learning through a deep neural network. In: Proceedings of the 28th ACM international conference on information and knowledge management, pp 1853–1862
179. Xu Y, Xu Y, Saria S (2016) A Bayesian nonparametric approach for estimating individualized treatment-response curves. In: Machine learning for healthcare conference, pp 282–300
180. Xu Y (2017) Generalized synthetic control method: causal inference with interactive fixed effects models. Polit Anal 25(1):57–76
181. Yoon J, Jarrett D, van der Schaar M (2020) Google chrome privacy whitepaper. In: Curran associates, Inc. http://papers.nips.cc/paper/8789-time-series-generative-adversarial-networks.pdf
182. Zhang K, Chan LW (2006) Extensions of ICA for causality discovery in the hong kong stock market. In: International conference on neural information processing. Springer, pp 400–409
183. Zheng X et al (2017) State space LSTM models with particle MCMC inference. arXiv:1711.11179 [cs.LG]
184. Zhu L, Lu W, Song R (2020) Causal effect estimation and optimal dose suggestions in mobile health. In: ICML



R. Moraffah received her B.S. degree in Computer Science and Engineering from Sharif University of Technology. She is currently a fifth-year PhD student of Computer Science and Engineering at Arizona State University. Her research interests include causal inference, causal ML and adversarial learning. Contact her at rmoraffa@asu.edu.

P. Sheth received his B.Tech degree in Computer Sience and Engineering from Institute of Engineering and Management. He is currently a third-year PhD student of Computer Science and Engineering at Arizona State University. His research interests include causal inference, data mining and causal ML. Contact him at psheth5@asu.edu.

M. Karami is a computer engineering Ph.D. student at Arizona State University working on various projects in the field of social media mining, machine learning, and natural language processing. She received her master's degree in computer engineering with a focus on artificial intelligence from Sharif University of Technology working on 3D reconstruction algorithms. Contact information: mkarami@asu.edu.



A. Bhattacharya received his B.Eng degree in Electronics and Communication Engineering from Birla Institute Of Technology. He is currently a PhD student at University Of Kentucky. His research interests include text generation, biomedical data science and 3D image reconstruction. Contact him at abh240@uky.edu

Q. Wang received the BE degree in software engineering from Northwestern Polytechnical University in 2016 and then received the master degree in computer science and technology from Northwestern Polytechnical University, 2018. She is currently pursuing the PhD degree at Northwestern Polytechnical University, Xi'an, China. Her current research interest includes urban computing and mobile crowd sensing. Contact her at qr369wang@gmail.com.

A. Tahir received his B.S. degree in Computer Science from Lahore University of Management Sciences. He received his M.S. degree in Computer Science from Arizona State University. He is currently a PhD student of Computer Science at Arizona State University. His research interests include Machine Learning and Data Mining. Contact him at artahir@asu.edu.

A. Raglin is a researcher at the U.S. Army Combat Capabilities Development Command, known as DEVCOM, Army Research Laboratory. Her research interests span causal inference, machine learning and uncertainty of information. Contact her at adrienne.raglin2.civ@mail.mil.

H. Liu is a professor of Computer Science and Engineering at Arizona State University. His research interests are in data mining, machine learning, social computing, and artificial intelligence. He is Editor in Chief of ACM TIST, and Field Chief Editor of Frontiers in Big Data and its Specialty Chief Editor of Data Mining and Management. He is a Fellow of ACM, AAAI, AAAS, and IEEE