A Practical Approach to Causal Inference over Time

Martina Cinquini[*1], **Isacco Beretta**[*1], **Salvatore Ruggieri**[1], **Isabel Valera**[2]

[1]Department of Computer Science, University of Pisa, Pisa, Italy
[2]Department of Computer Science, Saarland University, Saarbrücken, Germany
{martina.cinquini, isacco.beretta}@phd.unipi.it, salvatore.ruggieri@unipi.it, ivalera@cs.uni-saarland.de

Abstract

In this paper, we focus on estimating the causal effect of an intervention over time on a dynamical system. To that end, we formally define causal interventions and their effects over time on discrete-time stochastic processes (DSPs). Then, we show under which conditions the equilibrium states of a DSP, both before and after a causal intervention, can be captured by a structural causal model (SCM). With such an equivalence at hand, we provide an explicit mapping from vector autoregressive models (VARs), broadly applied in econometrics, to linear, but potentially cyclic and/or affected by unmeasured confounders, SCMs. The resulting *causal* VAR *framework* allows us to perform *causal inference over time* from observational time series data. Our experiments on synthetic and real-world datasets show that the proposed framework achieves strong performance in terms of observational forecasting while enabling accurate estimation of the causal effect of interventions on dynamical systems. We demonstrate, through a case study, the potential practical questions that can be addressed using the proposed causal VAR framework.

Code — https://github.com/marti5ini/ci-over-time
Extended version — https://arxiv.org/abs/2410.10502

1 Introduction

Dynamical systems often exhibit complex behaviors that unfold over time, leading to delayed responses and feedback loops. Understanding the causal effect of interventions within such systems is crucial across disciplines such as climate (Runge et al. 2019) and social sciences (Wunsch et al. 2022), where different time scales play a central role. For instance, monetary policy adjustments may have immediate effects on consumer spending, but their impact on inflation and economic growth only becomes evident in the medium/long-term. Similarly, the consequences of human actions on climate change may take decades to manifest, with the risk of endorsing public policies that underestimate their relevance. To address these issues, it is essential to perform *causal inference over time*. From the perspective of causality, structural causal models (SCMs) provide a formal framework to enable causal inference from cross-sectional data. However, adapting existing methods to capture temporal dynamics remains a challenge (Bongers et al. 2021). Alternatively, autoregressive models offer practical methods for time-series analysis and forecasting (Lütkepohl 2005), but their formalization of causal effects is limited. First, they model interventions as *shocks* in a specific point in time, with effects that fade away after a certain period (Moneta et al. 2011). Second, they rely on Granger causality (Granger 1969) which is concerned with how well one variable can predict another rather than identifying causal relations between them. Our work combines the strengths of both frameworks, i.e., SCMs and vector autoregressive models (VARs), to enable robust reasoning about the causal effect of interventions on dynamical systems over time. To that end, we first introduce a formal definition of causal interventions on discrete-time stochastic processes (DSPs), proposing two alternatives, additive and forcing interventions. Second, we establish conditions under which the equilibrium state of a DSP can be represented by an SCM. Third, we develop a framework that maps VARs to linear SCMs, handling potentially cyclic structures and unmeasured confounders. Finally, our framework is empirically validated on synthetic and real-world time-series data.

Related work The works most closely related to ours are these from Mooij, Janzing, and Schölkopf (2013) and Bongers, Blom, and Mooij (2018), as they theoretically connect dynamical systems to the causal semantics of SCMs via the equilibration of deterministic and random differential equations, and thus are capable of modeling cyclic causal mechanisms. Our approach differs from this line of work in two key aspects: i) we focus on *discrete-time* dynamical systems parameterized using *stochastic equations* which, as stated by Bongers, Blom, and Mooij (2018), become challenging for continuous-time processes; and ii) our mapping from autoregressive DSPs to SCMs provides not only a theoretical but also, to the best of our knowledge, *the first data-driven framework for performing causal inference over time in dynamical systems.*

2 Preliminaries and Background

2.1 Structural Causal Models

Equation A SCM $\mathcal{M} = (\mathbf{F}, \mathbf{E})$ determines how a set of d endogenous (observed) random variables $\mathbf{X} :=$

[*]These authors contributed equally.

$\{X^{(1)}, \dots, X^{(d)}\}$ are obtained from a set of exogenous variables $\mathbf{E} := \{E_1, \dots, E_d\}$, with prior distribution $p(\mathbf{E})$, via a set of structural equations $\mathbf{F} := \{X^{(i)} := f_i(\mathbf{PA}^{(i)}, \mathbf{E}^{(i)})\}_{i=1}^d$. Each f_i computes $X^{(i)}$ from its causal parents[1] $\mathbf{PA}^{(i)} \subseteq \boldsymbol{X}$ and a set $\mathbf{E}^{(i)} \subseteq \mathbf{E}$. We refer to $\boldsymbol{X}$ as a solution of $\mathcal{M}$. We assume $\mathbf{PA}^{(i)}$ to be minimal, i.e., it only contains variables $X^{(j)}$ such that $\partial_{X^{(j)}} f_i \neq 0$. This formulation extends the definition in Pearl (2009) to include cycles as in Bongers, Blom, and Mooij (2018).

Graph A SCM $\mathcal{M}$ induces a directed graph $\mathcal{G}_\mathcal{M} = (\mathcal{V}, \mathcal{E})$ that describes the functional dependencies in $\mathbf{F}$: $\mathcal{V}$ is the set of nodes for which V_i represents $X^{(i)}$ and $\mathcal{E}$ is the set of the edges $(V_i, V_j) \in \mathcal{E} \iff X^{(i)} \in \mathbf{PA}^{(j)}$.

Intervention Besides describing the observational distribution $p(\boldsymbol{X})$, SCMs allow answering *interventional queries* about the effect of external manipulations, and enable *counterfactual queries* assessing what would have happened to a particular observation if one observed variable $X^{(i)}$ had taken a different value. An intervention $\mathcal{I}$ on a SCM $\mathcal{M}$ yields a new SCM $\mathcal{M}^\mathcal{I}$ for which one or more mechanisms $f_i(\mathbf{PA}^{(i)}, \mathbf{E}^{(i)})$ change to $\tilde{f}_i(\tilde{\mathbf{PA}}^{(i)}, \tilde{\mathbf{E}}^{(i)})$, where $\tilde{\mathbf{PA}}^{(i)} \subseteq \mathbf{PA}^{(i)}$ and $\tilde{\mathbf{E}}^{(i)} \subseteq \mathbf{E}^{(i)}$. We refer to a *hard intervention* when f_i is replaced by a constant value $\alpha^{(i)}$, and $\tilde{\mathbf{PA}}^{(i)} = \tilde{\mathbf{E}}^{(i)} = \varnothing$. This type of intervention is denoted by the do-operator $do(X^{(i)} = \alpha^{(i)})$. On the other hand, we refer to a *soft intervention* when at least one argument of f_i is retained. The causal effect CE of an intervention is evaluated in terms of differences between the values of the observable variables before and after the intervention $\mathcal{I}$, i.e.,

$$\mathtt{CE}^\mathcal{I} = \mathbb{E}[\boldsymbol{X}^\mathcal{I} - \boldsymbol{X}]. \tag{1}$$

2.2 Discrete-Time Stochastic Processes

A discrete-time (vector) stochastic process (DSP) is a function $\boldsymbol{X} : T \times \Omega \to \mathbb{R}^d$ where $t \in T$ is a time index in $\mathbb{Z}$, such that $\boldsymbol{X}_t$ (denoting $\boldsymbol{X}(t, \cdot)$) is a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$[2]. We refer to $\boldsymbol{X}(\omega)$ (denoting $\boldsymbol{X}(\cdot, \omega)$) as a *realization* or *trajectory* of $\boldsymbol{X}$ and denote the i-th component of $\boldsymbol{X}$ with $X^{(i)}$. Every DSP can be described through a difference equation (dE), i.e., a recurrence relation that allows computing $\boldsymbol{X}_t$ based on its past values. dEs can be categorized into three types (Bongers, Blom, and Mooij 2018): ordinary difference equations (OdE) describing deterministic processes; random difference equations (RdE), which involve randomness in the initial state $\boldsymbol{X}_0$ and in the evolution parameters (see App. A); and stochastic difference equations, which describe inherently stochastic trajectories.

Equation A stochastic difference equation (SdE) describes a DSP via a functional relationship of the form

$$\boldsymbol{X}_t = \boldsymbol{f}(\boldsymbol{X}_{<t}) + \boldsymbol{g}(\boldsymbol{X}_{<t}) \odot \boldsymbol{\varepsilon}_t, \tag{2}$$

where $\boldsymbol{X}_{<t} := \{\boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-2}, \dots\}$ represents the trajectory up to time t, $\boldsymbol{f}$ represents the system's deterministic mechanism, and the Hadamard product $\boldsymbol{g} \odot \boldsymbol{\varepsilon}_t$ is the inhomogeneous stochastic part, where $\boldsymbol{\varepsilon}_t$ denotes *white noise*, i.e., $\forall t, t' \in T \quad \mathbb{E}[\boldsymbol{\varepsilon}_t] = \mathbf{0}, \mathbb{E}[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_{t'}] = \Sigma_{\boldsymbol{\varepsilon}}^2 \delta_{t,t'}$. While for OdEs and RdEs trajectories $\boldsymbol{X}(\omega)$ may asymptotically converge to an equilibrium, SdEs cannot exhibit such convergence due to the ongoing influence of $\boldsymbol{g} \odot \boldsymbol{\varepsilon}_t$.

Graph Analogously to SCMs, we can associate a directed graph $\mathcal{G}_\mathcal{D}$ to a dE $\mathcal{D}$, consisting of nodes V_i representing individual components $X^{(i)}$, while an edge (V_i, V_j) is present if $\exists k > 0$ such that $\partial_{X_t^{(i)}} X_{t+k}^{(j)} \neq 0$ in $\mathcal{D}$.

2.3 Vector Autoregressive Models

In this paper, we focus on a specific type of SdE, the VAR model (Kilian and Lütkepohl 2017).

Equation Consider a d-dimensional vector-valued stationary time series $\{\boldsymbol{X}_0, \dots, \boldsymbol{X}_T\}$ generated by a VAR model with *lag p*, where a lag represents the number of previous time steps used to predict the current value of each variable. Specifically, the VAR(p) model is defined by

$$\boldsymbol{X}_t = \boldsymbol{\nu} + \boldsymbol{A}_1 \boldsymbol{X}_{t-1} + \dots + \boldsymbol{A}_p \boldsymbol{X}_{t-p} + \boldsymbol{u}_t, \tag{3}$$

where $\boldsymbol{\nu}$ is a d-dimensional vector of intercept terms, $\{\boldsymbol{A}_i\}_{i=1}^p$ are $(d \times d)$ matrices and $\boldsymbol{u}_t$ is a d-dimensional white noise term. If the process $\boldsymbol{X}_t$ is stable and stationary (Hamilton 1994), Equation 3 can also be written as

$$\begin{aligned} &\boldsymbol{A}(L)\boldsymbol{X}_t = \boldsymbol{\nu} + \boldsymbol{u}_t, \\ \text{with} \quad &\boldsymbol{A}(L) := \boldsymbol{I}_d - \boldsymbol{A}_1 L - \dots - \boldsymbol{A}_p L^p, \end{aligned} \tag{4}$$

where L is the *lag operator* such that $L\boldsymbol{X}_t \equiv \boldsymbol{X}_{t-1}$ and $\boldsymbol{I}_d$ is a d-dimensional identity matrix.

A limitation of VARs is the inability to interpret the system in causal terms since the components of $\boldsymbol{u}_t$ are cross-correlated and act as hidden confounders. A common approach to overcome this issue is to orthogonalize the noise terms. In this context, the process of *causal discovery*, i.e., inferring the causal structure of the data, is analogous to the one of SCMs (Hyvärinen et al. 2010; Moneta et al. 2013; Geiger et al. 2015; Malinsky and Spirtes 2018), and identifies a triangular matrix $\hat{\boldsymbol{A}}_0$ such that $\boldsymbol{\varepsilon}_t = \hat{\boldsymbol{A}}_0 \boldsymbol{u}_t$ consists of mutually uncorrelated elements. The transformed VAR, known as the Structural VAR (SVAR) model in the literature (Kilian and Lütkepohl 2017), is defined by $\hat{\boldsymbol{A}}_0 \boldsymbol{X}_t = \hat{\boldsymbol{A}}_0 \boldsymbol{\nu} + \hat{\boldsymbol{A}}_1 \boldsymbol{X}_{t-1} + \dots + \hat{\boldsymbol{A}}_p \boldsymbol{X}_{t-p} + \boldsymbol{\varepsilon}_t$, where $\hat{\boldsymbol{A}}_i = \hat{\boldsymbol{A}}_0 \boldsymbol{A}_i$. From a modeling perspective, VAR and SVAR are equivalent, as any SVAR can be expressed in its reduced-form VAR by computing $\boldsymbol{A}_i = \hat{\boldsymbol{A}}_0^{-1} \hat{\boldsymbol{A}}_i$ for $i = 0, \dots, p$ in Eq. 3. Notably, choosing one over the other does not affect its causal interpretation, provided that $\hat{\boldsymbol{A}}_0$ is known. For simplicity, in this work, we adopt the VAR notation, to introduce a *novel framework for causal inference over time, which complements the* SVAR*'s causal discovery approach*.

Graph An edge (V_i, V_j) is present iff $\exists k$ s.t. $\boldsymbol{A}_k[i, j] \neq 0$.

[1]Unlike in acyclic SCMs, $\mathbf{PA}^{(i)}$ loses its hierarchical interpretation since two nodes can be mutually parents.

[2]The probability space $(\Omega, \mathcal{F}, \mathbb{P})$ consists of a sample space Ω, a σ-algebra $\mathcal{F}$ of events, and a probability measure $\mathbb{P}$ that assigns probabilities to events in $\mathcal{F}$.

3 Causal Perspective on Discrete-Time Stochastic Processes

This section provides the theoretical basis for causal inference over time. First, we formally define causal interventions on SdEs (§3.1). Then, we show how a SCM can be considered a compressed description of the asymptotic behavior of an underlying dynamical system (§3.2).

3.1 Causal Interventions on SdEs

We define an intervention $\mathcal{I}$ on a SdE $\mathcal{D}$ as a modification of one or more component equations denoted by the mapping:

$$\begin{aligned}\mathcal{I} : f_i(\mathbf{PA}^{(i)}_{<t}) + g_i(\mathbf{PA}^{(i)}_{<t}) \odot \boldsymbol{\varepsilon}_t \quad \longmapsto \\ \tilde{f}_i(\tilde{\mathbf{PA}}^{(i)}_{<t}) + \tilde{g}_i(\tilde{\mathbf{PA}}^{(i)}_{<t}) \odot \boldsymbol{\varepsilon}_t,\ \forall t \geq t_{\mathcal{I}}\end{aligned} \tag{5}$$

where $\tilde{\mathbf{PA}}^{(i)}_{<t} \subseteq \mathbf{PA}^{(i)}_{<t}$. Unlike SCMs, the intervention applies *starting from a specific time* $t_{\mathcal{I}}$. In other words, the process follows the original equations for $t < t_{\mathcal{I}}$ and the modified ones for $t \geq t_{\mathcal{I}}$. We denote the modified SdE as $\mathcal{D}^{\mathcal{I}}$ to generalize Eq. 1 to account for time. To differentiate between interventions on a SCM $\mathcal{M}$ and on a SdE $\mathcal{D}$, $\mathcal{I}_{\mathcal{M}}$ and $\mathcal{I}_{\mathcal{D}}$ will be respectively adopted when necessary.

Definition 1 (Causal Effect over time ($\mathtt{CE}_t$)). Let $\boldsymbol{X}$ be a solution of a specific SdE $\mathcal{D}$. We define the causal effect at time t of an intervention $\mathcal{I}$ as

$$\mathtt{CE}^{\mathcal{I}}_t := \mathbb{E}[\boldsymbol{X}^{\mathcal{I}}_t - \boldsymbol{X}_t | \boldsymbol{X}_{<t_{\mathcal{I}}}], \tag{6}$$

where $\boldsymbol{X}^{\mathcal{I}}$ is the solution of the modified SdE $\mathcal{D}^{\mathcal{I}}$.

The interpretation of $\mathtt{CE}^{\mathcal{I}}_t$ is closely related to the causal effect of an intervention on a SCM (Eq. 1), $\mathtt{CE}^{\mathcal{I}}$: while the latter measures the causal effect of an exogenous intervention, $\mathtt{CE}_t$ does so for any time step t of the DSP, i.e., *it measures the causal effect of an intervention over time*. Importantly, as we will show in the next section, $\mathtt{CE}^{\mathcal{I}}_t \to \mathtt{CE}^{\mathcal{I}}$ as $t \to \infty$, i.e., there is an asymptotic correspondence between the two quantities.

3.2 Mapping SdEs to SCMs

Given an SdE $\mathcal{D}$ and its solution $\boldsymbol{X}$, we study the conditions on $\mathcal{D}$ such that: i) $\boldsymbol{X}_t$ converges in distribution to $\boldsymbol{X}_\infty$ as $t \to \infty$; and ii) there exists an SCM $\mathcal{M}$ such that $\boldsymbol{X}_\infty$ is a solution of $\mathcal{M}$ and, for every intervention $\mathcal{I}$, it holds that $(\boldsymbol{X}_\infty)^{\mathcal{I}_{\mathcal{M}}} = (\boldsymbol{X}^{\mathcal{I}_{\mathcal{D}}})_\infty$. While i) is automatically satisfied by any finite memory stationary process, ii) requires more careful analysis, as discussed below.

A negative result from (Janzing, Rubenstein, and Schölkopf 2018) Consider the stable bivariate system defined by the equations $X_t = \varepsilon^x_t, Y_t = 0.5 \cdot X_{t-1} + \varepsilon^y_t$. For every t, X_t and Y_t are independent of each other. Consequently, the joint distribution $p(X_t, Y_t)$ cannot capture the causal dependencies of the system (X *causes* Y). The lack of causal information in the cross-sectional dimension arises because the variables are *localized in time*; their values change rapidly, leading to minimal or no correlation with their past values. On this specific point Janzing, Rubenstein, and Schölkopf (2018) provide an explicit negative result: without first making the variables de-localized in time, there is no SCM that can capture the SdE. In fact, our definition of intervention (Eq. 5) *acts on a variable of the system for a prolonged and indefinite period*.

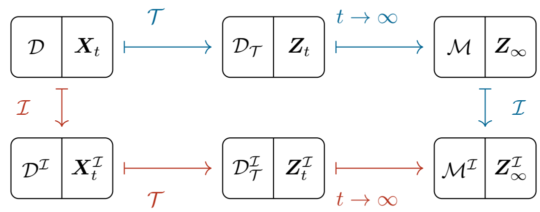


Figure 1: $\mathcal{T}$-transformation transfers causal information from the temporal to the cross-sectional dimension, and thus to the joint distribution $P(\boldsymbol{Z}_t)$. The diagram commutes, i.e., red and blue paths produce the same result.

$\mathcal{T}$-transformation To overcome this limitation, (Janzing, Rubenstein, and Schölkopf 2018) propose a transformation of $\boldsymbol{X}_t$ based on a frequency analysis of the time series.[3] Instead, our choice is inspired by the long-run normalized mean via the transformation $\mathcal{T}$: DSP $\mapsto$ DSP defined by

$$\mathcal{T}(\boldsymbol{X})_t := \boldsymbol{Z}_t = \boldsymbol{\mu} + \frac{1}{\sqrt{t}} \sum_{i=1}^{t} (\boldsymbol{X}_i - \boldsymbol{\mu}), \tag{7}$$

where $\boldsymbol{\mu} := \mathbb{E}[\boldsymbol{X}]$.[4] Moreover, $\mathbb{E}[\boldsymbol{Z}_\infty] = \mathbb{E}[\boldsymbol{X}_\infty] = \boldsymbol{\mu}$ so that for every intervention $\mathcal{I}$, $\mathtt{CE}^{\mathcal{I}}_\infty$ (Eq. 6) *yields the same values*. However, unlike $\boldsymbol{X}$, $\boldsymbol{Z}$ can be mapped into an SCM that precisely models *its distribution shift over any intervention*, thereby satisfying property ii), represented as the commutativity of the diagram in Fig. 1.

It is important to clarify that $\boldsymbol{Z}_t$ is not the process of interest, and the focus of the causal analysis remains on $\boldsymbol{X}_t$. However, due to the equivalence of long-run causal effects calculated in both processes, and the ability to associate $\boldsymbol{Z}_t$ with the SCM that models these effects, $\boldsymbol{Z}_t$ serves as a convenient intermediate mathematical tool. To demonstrate how this transformation ensures these desirable properties, we will focus on the subclass of linear systems, particularly on VAR models. The reason for this choice is twofold. First, linear models, despite their simplicity, are still on par performance-wise with state-of-the-art Machine-Learning based forecast techniques (Toner and Darlow 2024), in particular when dealing with stochastic time series (Parmezan, de Souza, and Batista 2019). Second, the mathematical treatment of interventions and the estimation of causal effects is particularly straightforward to implement and interpret, making this a useful first step for a possible extension to the nonlinear case.

[3]Our $\boldsymbol{Z}_t$ (Eq. 7) can also be interpreted as a form of discrete Fourier transform of the time-series $\boldsymbol{X}_{1:t}$ with frequency zero.

[4]The expectation here is taken over time as well. Nonetheless, for stationary processes, this simplifies to $\mathbb{E}[\boldsymbol{X}] = \mathbb{E}[\boldsymbol{X}_t]$ for all t.

4 From Vector Autoregressive Models to Structural Causal Models

In this section, we show that linear SCMs can model the long-term effects of stable VARs, explaining the properties of its DSP equilibrium (§4.1). Then, we provide implementations of two types of causal interventions, leveraging the strengths of the VAR framework (§4.2). Finally, we discuss the practical implications of our theoretical results (§4.3).

4.1 Mapping from VARs to SCMs

We provide the explicit mapping from VARs to linear SCMs in the following theorem (proved in App. B.2).

Theorem 1. Given a stable VAR(p) $\mathcal{D}$ defined by Eq. 3, there exists a linear SCM $\mathcal{M}$ with structural equations [5]

$$\begin{aligned} &\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{X}} + \tilde{\boldsymbol{u}}, \\ \text{where} \quad &\tilde{\boldsymbol{A}} := [\boldsymbol{A}_1 + \cdots + \boldsymbol{A}_p] \text{ and } \tilde{\boldsymbol{u}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}}), \end{aligned} \tag{8}$$

such that, given the transformation $\boldsymbol{Z}_t = \frac{1}{\sqrt{t}} \sum_{i=1}^{t} \boldsymbol{X}_i$, the following properties hold:

1. $\mathcal{D}$ and $\mathcal{M}$ share the same *causal graph*, i.e., $\mathcal{G}_{\mathcal{M}} = \mathcal{G}_{\mathcal{D}}$;
2. The *observational distribution* induced by $\mathcal{D}$ at equilibrium $p(\boldsymbol{Z}_\infty)$ is equal to the one induced by $\mathcal{M}$, $p(\tilde{\boldsymbol{X}})$;
3. The *interventional distribution* $p(\boldsymbol{Z}_\infty^{\mathcal{I}_\mathcal{D}})$ is equal to the one induced by the same intervention on $\mathcal{M}$, $p(\tilde{\boldsymbol{X}}^{\mathcal{I}_\mathcal{M}})$.

Remark. Note that due to the influence of time in VARs, the equivalent SCMs at equilibrium, while linear, may lead to cycles in the causal graph (see, e.g., Fig. 2c) and correlations between the exogenous variables, captured by the full covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ in Eq. 8. Note also that the above Theorem implies that there is a direct relationship between interventions on DSPs, $\mathcal{I}_\mathcal{D}$, and interventions on SCMs, here denoted by $\mathcal{I}_\mathcal{M}$. Refer to App. B.2 for further details.

4.2 Implementation of Causal Interventions

Different application scenarios may need different types of interventions. Consider a government's fiscal policy. In such a setting, a feasible approach would be to implement an *additive intervention* in the form of an annual tax increase of, e.g., 300 euros per household on top of existing taxes. In other scenarios, e.g., when studying the effect of the *key European Central Bank's interest rate* (Belke and Polleit 2007), a more natural choice is to implement a *forcing intervention* that enforces the convergence of an observed variable (e.g., interest rate) to a target value. In the following, we propose an implementation for VARs of these two forms of interventions, showing their effects on the system and discussing their stability conditions.

Additive Interventions Given a stable VAR(p) as in Eq. 4, we define an additive intervention $\mathcal{I}_a$ at time $t_\mathcal{I}$ with force $\boldsymbol{F}$ as the mapping:

$$\begin{aligned} \mathcal{I}_a : \boldsymbol{A}(L)\boldsymbol{X}_t = \boldsymbol{\nu} + \boldsymbol{u}_t \quad &\longmapsto \\ \boldsymbol{A}(L)\boldsymbol{X}_t = \boldsymbol{\nu} + \boldsymbol{u}_t + \mathbb{I}(t \geq t_\mathcal{I})\boldsymbol{F}&, \end{aligned} \tag{9}$$

[5]For simplicity we set $\boldsymbol{\nu} = 0$, i.e., we assume $\mathbb{E}[\boldsymbol{X}_t] = 0$. The theorem applies in the general case up to a translation of both the VAR and the associated SCM.

where $\mathbb{I}(t \geq t_\mathcal{I})$ is the indicator function, which equals 1 if $t \geq t_\mathcal{I}$, otherwise 0. In other words, we perform a translation while keeping the process dynamics unchanged. In such case, the temporal causal effect $\texttt{CE}_t$ is deterministic and takes values $\texttt{CE}_t = 0$ for $t < t_\mathcal{I}$ while, for $k \geq 0$:

$$\texttt{CE}_{t_\mathcal{I}+k} = \sum_{l=0}^{k} \Phi_l \boldsymbol{F},$$

where Φ is the *impulse response function* of the VAR model. Refer to App. B.1 for further details.

Remark. For this type of intervention, $\texttt{CE}_t$ is deterministic and does not depend on the specific trajectory. The same property can be observed on the linear SCM associated with the process, defining the intervention in a similar way: $\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{X}} + \tilde{\boldsymbol{u}}$ changes into $\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{X}} + \boldsymbol{F} + \tilde{\boldsymbol{u}}$.

Stability Additive intervention preserves the stability regardless of the value of $\boldsymbol{F}$, since $\boldsymbol{A}(L)$ does not change. See App. B.1 for stability conditions of VARs.

Forcing Interventions We define a forcing intervention $\mathcal{I}_f$ at time $t_\mathcal{I}$ with force $\boldsymbol{F}$ and target value $\hat{\boldsymbol{X}}$ as:

$$\begin{aligned} \mathcal{I}_f : \boldsymbol{A}(L)\boldsymbol{X}_t = \boldsymbol{\nu} + \boldsymbol{u}_t \quad &\longmapsto \\ \boldsymbol{A}(L)\boldsymbol{X}_t = \boldsymbol{\nu} + \boldsymbol{u}_t + \mathbb{I}(t \geq t_\mathcal{I})\boldsymbol{F} &\odot (\hat{\boldsymbol{X}} - \boldsymbol{X}_t). \end{aligned} \tag{10}$$

We assume $\boldsymbol{F}$ to be positive in each component. This intervention acts as an attraction towards $\hat{\boldsymbol{X}}$, and $\boldsymbol{F}$ modulates the intensity of the attraction force. Applying an intervention on a single component $X^{(i)}$ toward the fixed value $\hat{X}$ and letting $F^{(i)} \to +\infty$ yields the do operator $do(X^{(i)} = \hat{X})$. We refer to Mooij, Janzing, and Schölkopf (2013) for a detailed discussion of this point.

Stability Forcing interventions $\mathcal{I}_f$ perturb the system dynamics by modifying the operator $\boldsymbol{A}(L)$. Specifically, by shifting the term $\boldsymbol{F} \odot \boldsymbol{X}_t$ to the left of the equation and rewriting it in matrix form as $\boldsymbol{F}_{diag}\boldsymbol{X}_t$, we obtain $\tilde{\boldsymbol{A}}(L) := \boldsymbol{A}(L) + \boldsymbol{F}_{diag}$. Hence, the stability of the intervened system is not guaranteed (we provide an example in App. B.3), and it is necessary to verify *that all the eigenvalues of* $\tilde{\boldsymbol{A}}(L)$ *are still inside the unit circle.* Intuitively, the stability of an observational system often relies on negative feedback loops. Fixing one variable can disrupt this balance, leading to runaway behavior. For example, turning off a pressure release valve in a pressurized tank can cause the pressure to build up uncontrollably, eventually leading to an explosion.

4.3 Practical Implications

Causal queries Our formulation of causal interventions on VARs differs from the standard approach based on Granger causality by being closer to that of SCMs. Consequently, it enables the generalization of interventional and counterfactual queries to account for time (see App. C). That is, it allows for answering the following causal questions:

- **Forecasted Interventions** What are the expected effects on an individual trajectory (or a population) when intervening in the present, and how do they vary over time?

- **Retrospective Counterfactuals** What would have happened to an individual trajectory if an intervention had been applied at a specific point in the past? What state would it be in now?

Both causal queries acquire a meaning embedded in the temporal dimension in terms of *forecasting for the future* (§5.2) and *retrospection for the past* (App D.3), respectively.

Expressiveness and universality VARs, despite their linearity, possess a high level of expressiveness (Kilian and Lütkepohl 2017). In fact, the Wold decomposition Theorem (Wold 1938) implies that the dynamics of *any purely non-deterministic covariance-stationary process can be approximated arbitrarily well by an autoregressive model, making them universal approximators*. In practice, linear autoregressive models are broadly used in time-series analysis. Yet, we intend to explore non-linear DSPs in future work, as they may lead to better convergence rates and allow for causal interpretation of a broader family of dynamical models.

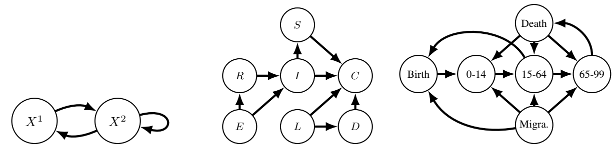
Feedback loops To properly understand complex systems, it is often useful to model feedback loops between their variables. Time-series models naturally capture this property, while SCMs require significant reformulation. The theory of cyclic SCMs has seen a significant advancement in recent years (Bongers et al. 2021), but practical approaches, both for causal discovery and causal inference, are still underdeveloped (Bongers et al. 2016; Lorbeer and Mohsen 2023). Our formalization of causal inference on VARs is a step forward in this direction.

Fitting VARs estimation is typically performed using ordinary least squares. Various alternative methods are available, both in terms of constrained optimization (e.g., to use prior knowledge about some coefficients of the VAR matrices (Sims 1980)) and within a Bayesian framework (Koop, Korobilis et al. 2010). Refer to Lütkepohl (2005, chapters 3,4,5) for a comprehensive discussion. Importantly, although VARs are most commonly used on time-series data (i.e., data from one single unit across a period of time), there are approaches tailored to the analysis of *panel data* (i.e., the evolution of many units over time) (Sigmund and Ferstl 2021); and *cross-sectional data* (i.e., many individuals at a single point of time), provided that they have at least some proxy variables of time (Deaton 1985). Such approaches open up a promising line of future work that can further generalize VARs applicability for causal reasoning over time.

5 Empirical Evaluation

In this section, we evaluate VAR models' accuracy and expressiveness in multivariate time series, focusing on two forecasting dimensions: observational (§5.1) and interventional (§5.2). Additional results and in-depth descriptions can be found in App. D.

Datasets We rely on two synthetic datasets, German[6] and Pendulum, and the real-world Census dataset[7]. German simulates a loan approval scenario with seven variables. Pendulum is a two-variable system where $X^{(1)}$ operates as a stabilizer for $X^{(2)}$, which exhibits a divergent dynamic. Census includes demographic variables across three age groups, along with migration, birth, and death rates from 1992 to 2023 for 50 countries. Fig. 2 illustrates the causal graphs for all datasets. See App. D for further details.



(a) Pendulum (b) German (c) Census

Figure 2: **Causal graphs**. The causal graph for (a) and (b) is known, while for (c), it is assumed. In (b), nodes are labeled with the initials of each feature: Expertise, Responsibility, Loan Amount, Duration, Income, Savings, and Credit Score. In (c), $0-14$, $15-64$, and $65-99$ represent age groups.

Metrics We measure the discrepancy between the h-steps forecast $\hat{\boldsymbol{X}}_{t+h}|\boldsymbol{X}_{<t}$ and the true value $\boldsymbol{X}_{t+h}$ on the test set $\boldsymbol{X}_{test}$. We report Mean Absolute Error (MAE) focusing on the target variables (i.e., *Credit Score* for German, $X^{(1)}$ for Pendulum, and age groups for Census). See App. D.1 for other metrics. All results are averaged over ten runs.

5.1 Observational Forecasting

Baselines We compare VAR with three relevant works: i) DLinear (Zeng et al. 2023), a decomposition-based linear model that separates trend and seasonal components; ii) TSMixer (Chen et al. 2023), a Multi-layer Perceptron (MLP) based model that focuses on mixing time and feature dimensions; iii) TiDE (Das et al. 2023), a MLP based encoder-decoder model. To assess the effectiveness of the forecasting methods, we introduce an observational oracle forecaster that has full knowledge about the true data generating process and produces the optimal predictor, i.e., $\hat{\boldsymbol{X}}_{t+h}|\boldsymbol{X}_{<t} = \mathbb{E}[\boldsymbol{X}_{t+h}|\boldsymbol{X}_{<t}]$.

How does the VAR performance compare with SOTA models for forecasting multivariate time series? The observational forecasting results in Table 1 show performance across varying data sizes (i.e., number of instances) and forecast horizons for all datasets. VAR emerges as the top-performing model, consistently matching or closely approaching Oracle's scores for all datasets. DLinear usually achieves predictive accuracy close to VAR for 1-step forecasts, presumably due to the common linear nature of both models. TiDE and TSMixer consistently underperform compared to VAR and DLinear for German and Pendulum. For all models (including Oracle), performance on the Pendulum dataset is uniformly worse than on the German, highlighting the greater challenge in forecasting given the system's stronger stochasticity and variables changing more rapidly over time. On Census, VAR and TiDE provide the best results, TiDE slightly outperforming VAR in 5-step horizon.

[6]This dataset is inspired on https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data

[7]https://data.census.gov/

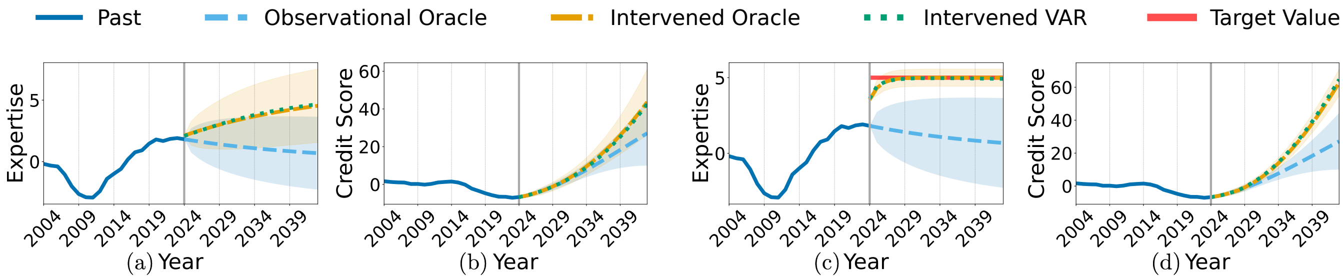


Figure 3: **Additive Intervention.** (a) Intervention on *Expertise* with $\boldsymbol{F} = 0.2$. (b) Effect on *Credit Score*, given the same intervention. **Forcing Intervention.** (c) Intervention on *Expertise* with $\boldsymbol{F} = 1$ and target $\hat{E} = 5$. (d) Effect on *Credit Score*. Shaded regions in plots denote 95% confidence bounds. Note that confidence bounds in (c)-(d) are narrower w.r.t. (a)-(b).

			Observational Forecasting				
Dataset	Size	Horizon	Oracle	VAR	DLinear	TiDE	TSMixer
German	100	1	`.004`	**.008**	.009	.011	.014
Pendulum	100	1	`.042`	**.043**	**.043**	.218	.217
German	100	10	`.014`	**.055**	**.055**	.094	.139
Pendulum	100	10	`.399`	**.420**	.440	1.43	1.43
German	500	1	`.004`	**.004**	**.004**	.011	.014
Pendulum	500	1	`.042`	**.042**	**.042**	.218	.217
German	500	10	`.014`	**.015**	**.015**	.093	.135
Pendulum	500	10	`.399`	**.401**	.405	1.43	1.43
Census	50×32	1	-	**.001**	.006	**.001**	.008
Census	50×32	5	-	.017	.025	**.014**	.024

Table 1: **Observational Forecasting.** MAE scores (*lower is better*) for VAR, DLinear (Zeng et al. 2023), TiDE (Das et al. 2023) and TSMixer (Chen et al. 2023). Results averaged over 10 runs. Due to space limitations, standard deviations are reported in App. D.1. Best model in bold, Oracle in `typewriter`. For Census, size equals the number of countries times the number of years.

			Interventional Forecasting	
Dataset	Size	Horizon	Additive	Forcing
German	100	1	$.000_{.000}$	$.000_{.000}$
German	100	10	$.043_{.028}$	$.364_{.297}$
German	500	1	$.000_{.000}$	$.000_{.000}$
German	500	10	$.018_{.014}$	$.115_{.081}$

Table 2: **Interventional Forecasting.** MAE scores (*lower is better*) for the proposed causal VAR framework on the German dataset. Results averaged over 10 runs, with standard deviation in subscript. Scores are scaled by a factor of 10^2 to ease readability.

5.2 Interventional Forecasting

We evaluate the causal VAR's forecast in estimating the causal effects on German. See App. D.2 for other datasets.

Baselines Since state-of-the-art methods do not allow computing the causal effect of interventions on dynamical systems, we use an oracle forecaster as a benchmark for theoretically optimal performance. Specifically, the ground truth values are estimated as $\texttt{CE}_{t+h} = \mathbb{E}[\boldsymbol{X}^{\mathcal{I}}_{t+h} - \boldsymbol{X}_{t+h} | \boldsymbol{X}_{<t}]$, while the predicted values from the proposed VAR framework as $\hat{\texttt{CE}}_{t+h} = (\hat{\boldsymbol{X}}^{\mathcal{I}}_t - \hat{\boldsymbol{X}}_{t+h}) | \boldsymbol{X}_{<t}$.

Interventions We perform causal interventions on the root node *Expertise* and observe the effect on the target variable *Credit Score*. For the additive case, we apply $\boldsymbol{F} = 0.2$, while for forcing, we use $\boldsymbol{F} = 1$ with a target value of $\hat{E} = 5$. These values are selected for illustrative purposes such that the long-term expected value of *Expertise* is the same for both interventions (i.e., 5). See App. D.2 for other variants.

How do additive and forcing interventions affect the system dynamics? Fig. 3a and Fig. 3b illustrate additive intervention. *Expertise* is a variable that typically necessitates several years for acquisition in practical scenarios. The causal VAR accurately captures such delayed impact as its effect on *Credit Score* appears after several years. As the system maintains its dynamic characteristics unchanged, the forecasted covariance remains the same even after the intervention. Fig. 3c and Fig. 3d show forcing intervention, where the interventional forecasting exactly aligns with the target value. Moreover, we stress that even for a low value of $\boldsymbol{F}$, the forcing intervention resembles a do-intervention (shrinking the variances significantly) even though theoretically convergence is guaranteed only for $F \to \infty$.

How accurate is the causal VAR framework in estimating the causal effect of interventions over time? Table 2 summarizes results on interventional forecasting, showing errors with varying data size and forecast horizons. At 1-step, both interventions lead to perfect performance since *Credit Score* is a slow-changing variable and requires at least 3 time steps for an intervention to take effect. At 10-step, our causal effect estimates remain highly accurate.

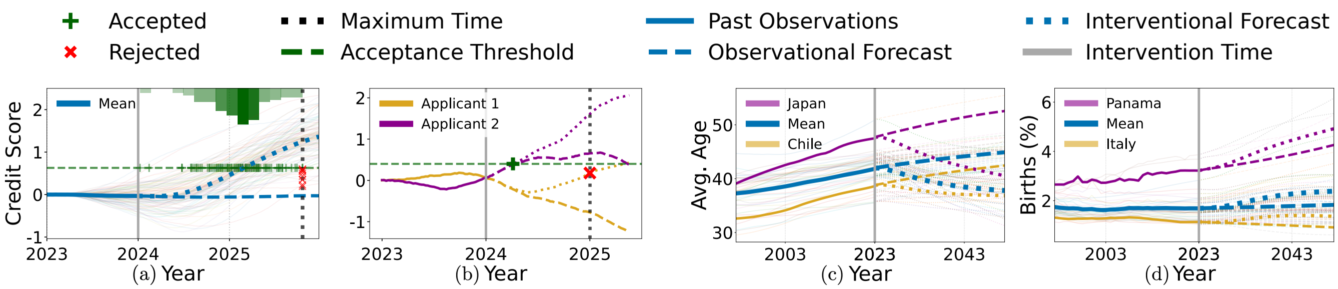


Figure 4: **German**: effect of increasing *Expertise* on *Credit Score*. (a) Time each loan applicant takes to cross or not the acceptance threshold. The histogram shows the crossing times distribution. (b) Comparison of two loan applicants with similar scores at intervention time. After the intervention, they diverge significantly, with only an applicant being accepted at the maximum time. Forecasts are dashed for observational and dotted for interventional. **Census**: additive intervention across all countries. (c) Intervention on *Births* with $\boldsymbol{F} = 0.004$ and its effect on the population's average age. (d) Intervention on *Migration* with $\boldsymbol{F} = 0.04$ and its effect on *Births*. In (c) and (d), we highlight two countries (one above and one below the mean) to illustrate the differences after the intervention. Forecasts are dashed for observational and dotted for interventional.

6 Use Cases

In this section, we show real-world scenarios where estimating causal effects over time represents a useful step toward a realistic modeling of the phenomena. We focus on the German and Census datasets, presenting two analyses for each. For German, Fig. 4a and Fig. 4b illustrate the effect of increasing *Expertise* by $\boldsymbol{F} = 0.38$ on *Credit Score*.

German 1 – Same intervention, different outcomes. Fig. 4a shows the distribution of the time required for loan applicants to cross or not the acceptance threshold after the intervention. Access to this information allows quantification of intervention efficacy, identification of credit-building patterns, and infer the key factors influencing loan eligibility outcomes. It can also inform the recommendation of actions (e.g. in algorithmic recourse (Karimi et al. 2023)) within a reasonable timeframe, fostering trust in the system and promoting user acceptance.

German 2 – Similar cross-sectional values, different causal effects over time. Fig. 4b shows trajectories that, while seeming similar at a given time, may have significantly divergent historical and future behaviors. For instance, the purple trajectory may have autonomously crossed the threshold without intervention, whereas in the case of the yellow one, the applied intervention may be inadequate to ensure the desired outcome. Such divergence highlights the importance of moving beyond models that rely only on cross-sectional data, motivating the need for techniques, such as the proposed causal VAR framework, that capture individual applicant behavior over time.

For Census, Fig. 4c and Fig. 4d present two additive interventions across all countries.

Census 1 – Impact of Births on Avg. Age. Fig. 4c shows the increase on *Births* with $\boldsymbol{F} = 0.004$ and its effect on the population's average age (computed as a weighted mean of age groups). The force value means that *Births* increase by 0.4% of each country's total population every year. Examining how they affect population age over time could allow policymakers to identify which countries might benefit most from specific types of demographic interventions and evaluate the long-term viability of systems. We can also observe that the intervention in Japan causes a more evident decrease in the average age than in Chile.

Census 2 – Impact of Migration on Births. Fig. 4d reports how a 4% growth in *Migration* w.r.t. the total population influences *Birth* rates. We observe that increased migration leads to a rise in births. However, its impact is less evident (observational and interventional forecasting trajectories are closer) compared to the result on the average age shown in Fig. 4c.

7 Concluding Remarks

In this work, we have established a link between discrete-time dynamical systems at equilibrium and SCMs. Moreover, we have provided an explicit procedure for mapping VARs to linear SCMs and demonstrated that, under specific model stability conditions, interventions on the dynamical system and the SCM yield equivalent results. To conduct causal inference over time, we have introduced two classes of interventions (additive and forcing) for VARs.

Limitations When systems exhibit strongly nonlinear dynamics, linear VARs may prove less effective than alternative nonlinear approaches. Moreover, our framework requires prior knowledge of the causal graph. In scenarios where this information is lacking, the process of causal discovery can present significant challenges.

Future work We will investigate the use of non-linear DSPs, as they may lead to better convergence rates and allow for causal interpretation of a broader family of dynamical models. Moreover, our work opens several interesting research directions (§4.3 for concrete examples) and applications (e.g. causal inference over time in high-dimensional contexts such as climate science).

Acknowledgments

This work has been partially supported by M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI"; and by the Italian Project Fondo Italiano per la Scienza FIS00001966 MIMOSA; and by the European Union (ERC-2018-ADG, XAI project, g.a. 834756); and by the NextGeneration EU programme - National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) for the project: "SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics" - Prot. IR0000013 - Avviso n. 3264 del 28/12/2021; and by the Deutsche Forschungsgemeinschaft (DFG. grant number 389792660 as part of the Transregional Collaborative Research Centre TRR 248: Center for Perspicuous Computing, CPEC); and by the European Union (ERC-2021-STG, SAML project, g.a. 101040177).

References

Belke, A.; and Polleit, T. 2007. How the ECB and the US Fed set interest rates. *Applied Economics*, 39(17): 2197–2209.

Bongers, S.; Blom, T.; and Mooij, J. M. 2018. Causal modeling of dynamical systems. *arXiv preprint arXiv:1803.08784*.

Bongers, S.; Forré, P.; Peters, J.; and Mooij, J. M. 2021. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5): 2885–2915.

Bongers, S.; Peters, J.; Schölkopf, B.; and Mooij, J. M. 2016. Structural Causal Models: Cycles, Marginalizations, Exogenous Reparametrizations and Reductions. *CoRR*, abs/1611.06221.

Chen, S.; Li, C.; Arik, S. Ö.; Yoder, N. C.; and Pfister, T. 2023. TSMixer: An All-MLP Architecture for Time Series Forecast-ing. *Trans. Mach. Learn. Res.*, 2023.

Das, A.; Kong, W.; Leach, A.; Mathur, S.; Sen, R.; and Yu, R. 2023. Long-term Forecasting with TiDE: Time-series Dense Encoder. *Trans. Mach. Learn. Res.*, 2023.

Deaton, A. 1985. Panel data from time series of cross-sections. *Journal of Econometrics*, 30(1-2): 109–126.

Geiger, P.; Zhang, K.; Schölkopf, B.; Gong, M.; and Janzing, D. 2015. Causal Inference by Identification of Vector Autoregressive Processes with Hidden Components. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, 1917–1925. JMLR.org.

Granger, C. W. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, 424–438.

Hamilton, J. D. 1994. Time Series Analysis.

Hyvärinen, A.; Zhang, K.; Shimizu, S.; and Hoyer, P. O. 2010. Estimation of a Structural Vector Autoregression Model Using Non-Gaussianity. *J. Mach. Learn. Res.*, 11: 1709–1731.

Janzing, D.; Rubenstein, P.; and Schölkopf, B. 2018. Structural causal models for macro-variables in time-series. *arXiv preprint arXiv:1804.03911*.

Karimi, A.; Barthe, G.; Schölkopf, B.; and Valera, I. 2023. A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *ACM Comput. Surv.*, 55(5): 95:1–95:29.

Kilian, L.; and Lütkepohl, H. 2017. *Structural vector autoregressive analysis*. Cambridge University Press.

Koop, G.; Korobilis, D.; et al. 2010. Bayesian multivariate time series methods for empirical macroeconomics. *Foundations and Trends® in Econometrics*, 3(4): 267–358.

Lorbeer, B.; and Mohsen, M. 2023. Comparative Study of Causal Discovery Methods for Cyclic Models with Hidden Confounders. In *CogMI*, 103–111. IEEE.

Lütkepohl, H. 2005. *New introduction to multiple time series analysis*. Springer Science & Business Media.

Malinsky, D.; and Spirtes, P. 2018. Causal Structure Learning from Multivariate Time Series in Settings with Unmeasured Confounding. In *CD@KDD*, volume 92 of *Proceedings of Machine Learning Research*, 23–47. PMLR.

Moneta, A.; Chlass, N.; Entner, D.; and Hoyer, P. O. 2011. Causal Search in Structural Vector Autoregressive Models. In *NIPS Mini-Symposium on Causality in Time Series*, volume 12 of *JMLR Proceedings*, 95–114. JMLR.org.

Moneta, A.; Entner, D.; Hoyer, P. O.; and Coad, A. 2013. Causal inference by independent component analysis: Theory and applications. *Oxford Bulletin of Economics and Statistics*, 75(5): 705–730.

Mooij, J. M.; Janzing, D.; and Schölkopf, B. 2013. From Ordinary Differential Equations to Structural Causal Models: the deterministic case. In *UAI*. AUAI Press.

Parmezan, A. R. S.; de Souza, V. M. A.; and Batista, G. E. A. P. A. 2019. Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model. *Inf. Sci.*, 484: 302–337.

Pearl, J. 2009. *Causality*. Cambridge university press.

Runge, J.; Bathiany, S.; Bollt, E.; Camps-Valls, G.; Coumou, D.; Deyle, E.; Glymour, C.; Kretschmer, M.; Mahecha, M. D.; Muñoz-Marí, J.; et al. 2019. Inferring causation from time series in Earth system sciences. *Nature Communications*, 10(1): 2553.

Sigmund, M.; and Ferstl, R. 2021. Panel vector autoregression in R with the package panelvar. *The Quarterly Review of Economics and Finance*, 80: 693–720.

Sims, C. A. 1980. Macroeconomics and reality. *Econometrica: Journal of the Econometric Society*, 1–48.

Toner, W.; and Darlow, L. N. 2024. An Analysis of Linear Time Series Forecasting Models. In *ICML*. OpenReview.net.

Wold, H. 1938. *A study in the analysis of stationary time series*. Ph.D. thesis, Almqvist & Wiksell.

Wunsch, G.; Russo, F.; Mouchart, M.; and Orsi, R. 2022. Time and causality in the social sciences. *Time & Society*, 31(2): 177–204.

Zeng, A.; Chen, M.; Zhang, L.; and Xu, Q. 2023. Are Transformers Effective for Time Series Forecasting? In *AAAI*, 11121–11128. AAAI Press.